Estadística

Ángel A. Juan
Blanca de la Fuente
Alicia Vila

PID_00159944

Material docente de la UOC

UOC
Universitat Oberta de Catalunya
www.uoc.edu

FONDO SOCIAL EUROPEO

MINISTERIO DE INDUSTRIA, TURISMO Y COMERCIO

plan avanza

alquimia

Ángel A. Juan

Licenciado en Matemáticas por la Universidad de Valencia, Máster en Tecnologías de la Información por la UOC y Doctor en Matemática Computacional Aplicada por la UNED. En la actualidad es profesor agregado de Estadística y Simulación en los Estudios de Informática, Multimedia y Telecomunicación de la UOC. Asimismo, es profesor asociado de Estadística Aplicada en la Universidad Politécnica de Cataluña. Sus líneas de investigación se centran en los ámbitos de la simulación por computador, el análisis de datos y el aprendizaje de las matemáticas en entornos en línea, ámbitos en los que ha publicado numerosos artículos en revistas y libros internacionales. Para más información, podéis consultar http://ajuanp.wordpress.com

Blanca de la Fuente

Doctora en Ciencias Biológicas (1988) por la Universidad Complutense de Madrid desde 1988. Profesora del Departamento de Estadística e Investigación Operativa II (Métodos de Decisión) de la Facultad de Ciencias Económicas y Empresariales de la Universidad Complutense de Madrid y Consultora de la Universitat Oberta de Catalunya. Ha sido docente desde 1992 de asignaturas del área de la Estadística en diversas titulaciones de universidades públicas y privadas. Sus áreas de investigación son el análisis multivariante y aplicaciones de nuevas metodologías docentes en la enseñanza universitaria.

Alicia Vila

Licenciada en Matemáticas por la Universidad de Valencia. Profesora de ciclos formativos en el ámbito de la informática, en particular en los campos de programación y bases de datos. Ha impartido docencia en el área de Probabilidad y Estadística en diferentes titulaciones de la Universitat Oberta de Catalunya.

El encargo y la creación de este material docente han sido coordinados por el profesor: Víctor Cavaller (2011)

El proyecto E-ALQUIMIA ha sido apoyado por el Ministerio de Industria, Turismo y Comercio en el marco de las ayudas para la realización de actuaciones sobre contenidos digitales en el marco del Plan Avanza, y por la Unión Europea a través de los Fondos Comunitarios. Referencia: PAV-10000-2007-275

Primera edición: febrero 2011
© Ángel A. Juan, Blanca de la Fuente y Alicia Vila
Todos los derechos reservados
© de esta edición, FUOC, 2011
Av. Tibidabo, 39-43, 08035 Barcelona
Realización editorial: Eureca Media, SL
Diseño: Manel Andreu
Depósito legal: B-1.339-2010
ISBN: 978-84-693-9717-6

Licencia Creative Commons, versión 3.0, modalidad BY-SA (attribution - share alike), que permite modificar la obra, reproducirla, distribuirla o comunicarla públicamente siempre que se reconozca su autoría y siempre que la obra derivada quede sujeta a la misma licencia que el material original.

Introducción

La asignatura de *Estadística* está dirigida a los estudiantes del grado de Información y Documentación.

Los estudios de Información y Documentación ofrecen múltiples salidas profesionales desde el trabajo en centros de información (bibliotecas, mediatecas, centros de documentación, archivos), hasta la gestión de información en organizaciones del sector privado o público (análisis de la información, gestión documental, gestión de contenidos, arquitectura de la información, webmaster) y la gestión de sistemas de información.

En general, la estadística se ha convertido en una herramienta imprescindible en el campo de las ciencias sociales, en los trabajos de investigación y a la hora de desarrollar profesionalmente tareas relacionadas con la gestión, la interpretación de datos y la toma de decisiones.

En el marco concreto de las competencias que tiene que desarrollar un gestor de la información y de la documentación, la estadística es un instrumento muy útil, sea cual sea el campo profesional que se quiere desarrollar.

Estos materiales introducen los conceptos estadísticos más necesarios para su formación, utilizando un enfoque práctico y aplicado. En este sentido, se da prioridad a la adquisición de conceptos y métodos aplicados, evitando el uso de un excesivo formalismo matemático. A priori, no se necesitan conocimientos previos de estadística, ya que esta asignatura se tratará desde cero y suponiendo que el estudiante no ha trabajado nunca en este campo.

El material didáctico está constituido por cinco módulos:

1. Estadística descriptiva, que incluye una introducción a la estadística y a la descripción de datos mediante tablas, gráficos y estadísticos, así como al concepto de probabilidad y de distribución de probabilidad.

2. Inferencia de información para una población, que incluye distribuciones, intervalos y contrastes.

3. Inferencia de información para dos poblaciones, sobre los contrastes de hipótesis para dos poblaciones.

4. Relación entre variables: causalidad, correlación y regresión, que incluye modelos de regresión simple (lineales, cuadráticos y cúbicos).

5. Introducción al diseño y análisis de encuestas, sobre las aplicaciones estadísticas a la selección de muestras y al análisis de cuestionarios.

Objetivos

El objetivo fundamental es introducir al estudiante en el uso de la metodología estadística para describir y compilar datos, construir muestras aleatorias válidas, comprobar hipótesis y elaborar modelos estadísticos.

A grandes rasgos, las competencias que se pretenden alcanzar son:

1. Entender la importancia de la estadística en la sociedad moderna.

2. Aprender a organizar y resumir de forma descriptiva un conjunto de datos de una muestra mediante gráficos, tablas de frecuencias y estadísticos.

3. Comprender el concepto de probabilidad de un acontecimiento y descubrir sus principales propiedades y aplicaciones.

4. Conocer las principales distribuciones estadísticas que se usan para modelar el comportamiento de variables discretas y continuas, y utilizarlas en pruebas de hipótesis.

5. Aplicar e interpretar la inferencia estadística en poblaciones.

6. Entender la importancia de las encuestas y los cuestionarios en la sociedad de la información y conocer su elaboración y aplicación.

7. Aprender a usar software estadístico y de análisis de datos como instrumento básico en la aplicación práctica de los conceptos y las técnicas estadísticas.

Contenidos

Bibliografía

Anderson, D.; Sweeney, D.; Williams, T. (2008). *Statistics for Business and Economics*. South-Western College Pub. ISBN: 0324658370.

Berk, K.; Carey, P. (2003). *Data Analysis with Microsoft Excel*. Duxbury Press. ISBN: 0534407145.

Bowermann, B. L.; O'Connell, R. T. (1997). *Applied Statistics: Improving Business Processes*. Irwin. ISBN: 025819386X.

Draper, N. R.; Smith, H. (1998). *Applied Regression Analysis. Wiley*. ISBN: 0471170828.

Fowler, F. (2008). *Survey Research Methods*. Sage Publications, Inc. ISBN: 1412958415.

Johnson, R.; Kuby, P. (2006). *Elementary Statistics*. Duxbury Press. ISBN: 0495017639.

Lohr, S. (1999). *Sampling: Design and Analysis*. Duxbury Press. ISBN: 0534353614.

Moore, D. (2006). *The Basic Practice of Statistics*. W. H. Freeman. ISBN: 071677478X.

Moore, D.; McCabe, G. (2005). *Introduction to the Practice of Statistics*. W. H. Freeman. ISBN: 0716764008.

Myer, R. H. (1990). *Classical and Modern Regression with Applications*. PWS. ISBN: 0534921787.

Rea, L.; Parker, R. (2005). *Designing and Conducting Survey Research: A Comprehensive Guide*. Jossey Bass. ISBN: 078797546X.

Ryan, B.; Joiner, B.; Cryer, J. (2005). *MINITAB Handbook. Brooks/Cole - Thomson Learning Inc*. ISBN: 0534496008.

Settle, R.; Alreck, P. (2003). *Survey Research Handbook*. McGraw-Hill/Irwin. ISBN: 0072945486.

Thompson, S. (2002). *Sampling*. Wiley-Interscience. ISBN: 0471291161.

Estadística descriptiva univariante

Modelos estadísticos para la descripción de datos univariantes

Alicia Vila y Ángel A. Juan

PID_00161058

Universitat Oberta de Catalunya
www.uoc.edu

Índice

Introducción

Las sociedades modernas son ricas en datos: la prensa escrita, la televisión y la radio, Internet y las intranets de las organizaciones ofrecen cantidades inmensas de datos que pueden ser procesados y analizados. Esto convierte a la estadística en una ciencia interesante y útil puesto que proporciona estrategias y herramientas que permiten obtener información a partir de dichos datos. Además, gracias a la evolución de la tecnología (ordenadores y software estadístico) hoy en día es posible automatizar gran parte de los cálculos matemáticos asociados al uso de técnicas estadísticas, lo que permite extender su uso a un gran rango de profesionales en ámbitos tan diversos como la biología, las ciencias empresariales, la sociología o las ciencias de la información.

La práctica de la estadística requiere aprender a obtener y explorar los datos –tanto numéricamente como mediante gráficos–, a pensar sobre el contexto de los datos y el diseño del estudio que los ha generado, a considerar la posible influencia de observaciones anómalas en los resultados obtenidos, a discutir la legitimidad de los supuestos requeridos por cada técnica y, finalmente, a validar la fiabilidad de las conclusiones derivadas del análisis. La estadística requiere tanto de conocimientos sobre los conceptos y técnicas empleados como de la suficiente capacidad crítica que permita evaluar la conveniencia de usar unas u otras técnicas según el tipo de datos disponible y el tipo de información que se desea obtener.

En este módulo inicial de la asignatura, se examinan los datos procedentes de una única variable: en primer lugar se explica cómo organizar y resumir dichos datos, tanto numérica como gráficamente (estadística descriptiva); en segundo lugar, se introducen los conceptos básicos asociados con la idea de probabilidad; finalmente, se presentan algunos modelos matemáticos que permiten analizar el comportamiento de algunas variables.

Objetivos

Los objetivos académicos que se plantean en este módulo son los siguientes:

1. Entender la importancia de la estadística en la sociedad moderna.

2. Aprender a organizar y resumir un conjunto de datos procedentes de una variable mediante gráficos, tablas de frecuencias y estadísticos descriptivos.

3. Comprender el concepto de probabilidad de un suceso y descubrir sus principales propiedades y aplicaciones.

4. Conocer las principales distribuciones estadísticas que se usan para modelar el comportamiento de variables discretas y continuas.

5. Saber calcular probabilidades asociadas a cada una de las distribuciones introducidas.

6. Aprender a usar software estadístico o de análisis de datos como instrumento básico en la aplicación práctica de los conceptos y técnicas estadísticas.

1. Introducción a la Estadística

La Estadística es la ciencia que se ocupa de obtener datos y procesarlos para transformarlos en información. Es, por tanto, un lenguaje universal ampliamente utilizado en las ciencias sociales, en las ciencias experimentales, en las ciencias de la salud y en las ingenierías. Las Tecnologías de la Información y la Comunicación (TIC) han incrementado notablemente la producción, diseminación y tratamiento de la información estadística. En particular, Internet es una fuente inagotable de datos que pueden ofrecer información y, a partir de ella, conocimiento. Por otra parte, la constante evolución de los ordenadores personales y de los **programas informáticos de estadística** y análisis de datos posibilita y facilita el análisis de grandes cantidades de datos mediante el uso de técnicas estadísticas y de minería de datos. En la Sociedad de la Información se hace pues imprescindible disponer de un cierto conocimiento estadístico incluso para poder comprender e interpretar correctamente los indicadores económicos (IPC, inflación, tasa de desempleo, Euribor, etc.), los indicadores bibliométricos (factor de impacto de una revista, cuartil en el que se sitúa, vida media de las citas recibidas, etc.) o los indicadores sociales (esperanza de vida, índice de alfabetización, índice de pobreza, indicador social de desarrollo sostenible, etc.) a los que frecuentemente se hace referencia en los medios de comunicación.

Nota

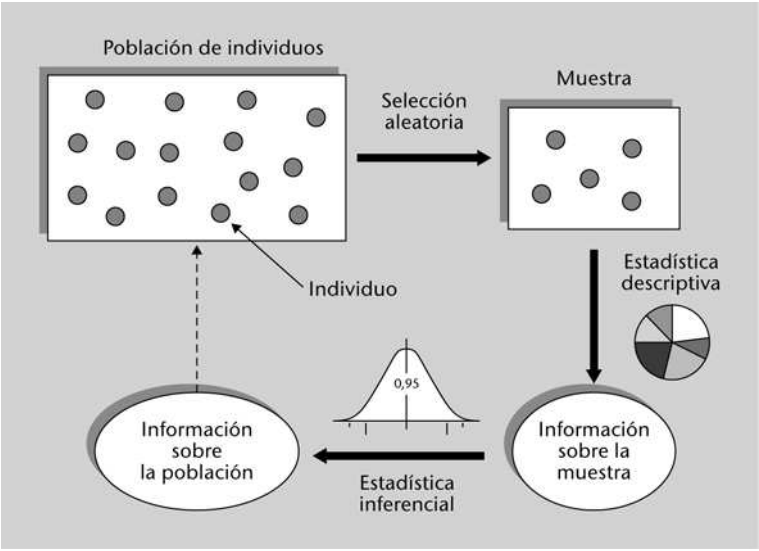
Las agencias gubernamentales, como el Instituto Nacional de Estadística (INE) o el Eurostat proporcionan datos sobre casi cualquier ámbito socioeconómico.

Software estadístico

En la actualidad existen excelentes **programas informáticos** para el análisis estadístico de datos. Algunos ejemplos son: MINITAB, SPSS, MS Excel, SAS, R, S-Plus, Statgraphics o Statistica.

El campo de la Estadística se puede dividir en dos grandes áreas: la estadística descriptiva y la estadística inferencial (figura 1).

Figura 1. Estadística descriptiva y estadística inferencial



La estadística descriptiva se ocupa de la obtención, presentación y descripción de datos procedentes de una muestra o subconjunto de una población de individuos. Por su parte, la estadística inferencial usa los resultados obtenidos

mediante la aplicación de las técnicas descriptivas a una muestra para inferir información sobre el total de la población a la que pertenece dicha muestra.

Algunos términos básicos

A lo largo de este material se usarán abundantes términos estadísticos, muchos de ellos bastante conocidos. A continuación se presentan y revisan algunos de estos términos básicos que conviene entender bien:

- **Población:** colección o conjunto de elementos (individuos, objetos o sucesos) cuyas propiedades se desean analizar. Ejemplos: (a) los estudiantes universitarios de un país; (b) el conjunto de periódicos en Internet; (c) el conjunto de revistas indexadas en el Science Citation Index (SCI), etc.

- **Muestra:** cualquier subconjunto de elementos de la población. Ejemplos: (a) los estudiantes de una determinada universidad; (b) los periódicos en línea centrados en aspectos económicos; (c) las revistas indexadas en el SCI de una determinada editorial, etc.

- **Muestra aleatoria:** muestra cuyos elementos han sido escogidos de forma aleatoria. Ejemplos: (a) un subconjunto de doscientos estudiantes escogidos al azar (mediante el uso de números aleatorios) de entre todos los matriculados en universidades de un país; (b) un subconjunto de cincuenta periódicos en línea escogidos al azar; (c) un subconjunto de quince revistas indexadas en el SCI escogidas al azar, etc.

- **Marco del muestreo:** lista que contiene aquellos elementos de la población candidatos a ser seleccionados en la fase de muestreo. No necesariamente coincidirá con toda la población de interés, ya que en ocasiones no será posible identificar a todos los elementos de la población. Ejemplos: (a) lista de todos los estudiantes matriculados en universidades de un país en un semestre concreto; (b) relación de periódicos en línea disponibles en un momento dado; (c) lista de todas las revistas indexadas en el SCI en un año específico, etc.

- **Variable aleatoria:** característica de interés asociada a cada uno de los elementos de la población o muestra considerada. Ejemplos: (a) la edad de cada estudiante; (b) el número de visitas diarias que recibe cada periódico en línea; (c) el factor de impacto de cada revista, etc.

- **Datos u observaciones:** conjunto de valores obtenidos para la variable de interés en cada uno de los elementos de la muestra. Ejemplos: (a) las edades registradas son {25, 23, 19, 28…}; (b) las visitas diarias registradas son {1326, 1792, 578, 982…}; (c) los factores de impacto registrados son {2,3; 1,7; 8,2…}.

- **Experimento:** estudio en la que el investigador controla o modifica expresamente las condiciones del mismo con la finalidad de analizar los distin-

tos patrones de respuesta en las observaciones. Ejemplos: (a) estudiar cómo varían las calificaciones de un grupo de estudiantes según dispongan o no de ordenadores con acceso a Internet en las aulas; (b) estudiar cómo varía el número de visitas a un periódico en línea según se opte o no por incluir noticias sensacionalistas en su portada; (c) estudiar cómo varía el factor de impacto de un grupo de revistas según éstas se incluyan o no en una base de datos de reconocido prestigio, etc.

- **Inspección o encuesta:** estudio en el que el investigador no pretende modificar las condiciones de la muestra con respecto a la variable de interés sino simplemente obtener los datos correspondientes a unas condiciones estándar. Ejemplos: (a) registrar las calificaciones de los estudiantes de un máster determinado; (b) realizar una encuesta a los lectores de un periódico en línea; (c) obtener el factor de impacto asociado a cada una de las revistas de una muestra, etc.

- **Parámetro:** valor numérico que sintetiza alguna propiedad determinada de la población. Los parámetros se asocian a toda la población y suelen representarse con letras del alfabeto griego como μ (mu), σ (sigma), etc. Ejemplos: (a) la edad media de todos los estudiantes universitarios de un país; (b) el número máximo de visitas diarias recibido por algún periódico en línea; (c) el rango o diferencia entre el mayor y el menor factor de impacto del conjunto de revistas indexadas en el SCI, etc.

- **Estadístico:** valor numérico que sintetiza alguna propiedad determinada de una muestra. Los estadísticos se asocian a una muestra y se suelen representar por letras del alfabeto latino como $\overline{x}$, s, etc. Ejemplos: (a) la edad media de los estudiantes de una muestra aleatoria; (b) el número máximo de visitas diarias recibidas por algún periódico deportivo en línea; (c) el rango o diferencia entre el mayor y el menor factor de impacto de las revistas de una editorial, etc.

- **Variable cualitativa o categórica:** variable que categoriza o describe cualitativamente un elemento de la población. Suele ser de tipo alfanumérico, pero incluso en el caso en que sea numérica no tiene sentido usarla en operaciones aritméticas. Ejemplos: (a) el teléfono o el correo electrónico de un estudiante; (b) la dirección IP de un periódico en línea; (c) el ISSN de una revista, etc.

- **Variable cuantitativa o numérica:** variable que cuantifica alguna propiedad de un elemento de la población. Es posible realizar operaciones aritméticas con ella. Ejemplos: (a) el importe de la beca que recibe un estudiante; (b) los ingresos que genera un periódico en línea; (c) el número de revistas publicadas por una editorial, etc.

- **Variable cuantitativa discreta:** variable cuantitativa que puede tomar un número finito o contable de valores distintos. Ejemplos: (a) edad de un es-

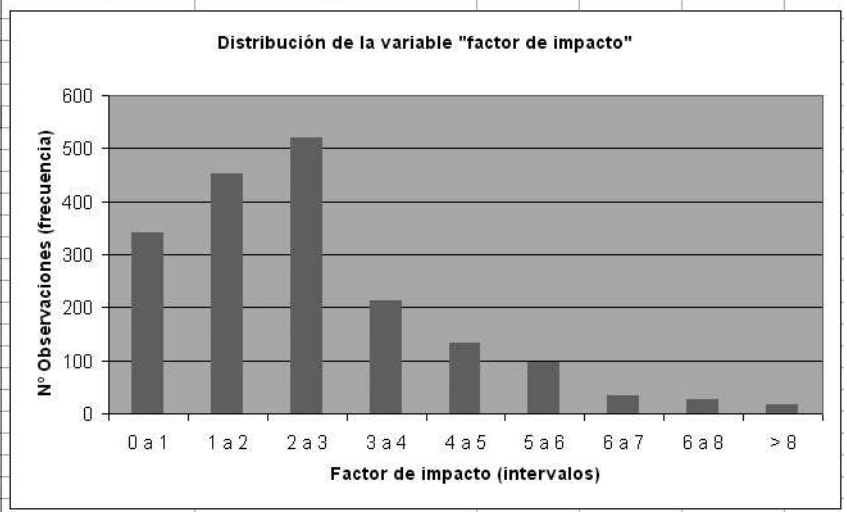
tudiante; (b) número de enlaces a otras fuentes de información que ofrece un periódico en línea; (c) calificación que obtiene una revista en una escala entera de 1 a 5, etc.

- **Variable cuantitativa continua:** variable cuantitativa que puede tomar un número infinito (no contable) de valores distintos. Ejemplos: (a) altura o peso de un estudiante; (b) tiempo que transcurre entre la publicación de una encuesta en línea y el instante en que ya la han completado un centenar de internautas; (c) factor de impacto (sin redondear) de una revista, etc.

- **Distribución de una variable:** en sentido amplio, una distribución es una tabla, gráfico o función matemática que explica cómo se comportan o distribuyen los valores de una variable, es decir, qué valores toma la variable así como la frecuencia de aparición de cada uno de ellos. Ejemplo: dada una muestra aleatoria de revistas, la distribución de la variable "factor de impacto de una revista" puede representarse mediante una tabla de frecuencias o mediante una gráfica como se aprecia en la figura 2. Se observa que trescientas cuarenta y dos de las revistas consideradas tienen un factor de impacto entre 0 y 1, cuatrocientas cincuenta y dos de las revistas tienen un factor de impacto entre 1 y 2, etc.

Figura 2. Distribución de una variable aleatoria

Factor de impacto	N.º de observaciones (frecuencia)
0 a 1	342
1 a 2	452
2 a 3	521
3 a 4	212
4 a 5	132
5 a 6	97
6 a 7	35
6 a 8	27
> 8	16



2. Descripción de datos mediante tablas y gráficos

Cuando se dispone de un conjunto de observaciones procedentes de una muestra conviene hacer un primer análisis exploratorio de éstas mediante gráficos y tablas que ayuden a interpretar los datos y a extraer información de los mismos. Existen diferentes tipos de gráficos que pueden usarse en esta fase exploratoria y el uso de unos u otros dependerá en gran medida del tipo de datos de los que se disponga (cualitativos o cuantitativos), así como de la información que se desee visualizar. En este apartado se presentaran algunos de los gráficos y tablas más habituales para la descripción de **datos univariantes**.

Datos univariantes

Los datos univariantes son los que provienen de una única variable. En algunos casos, los datos pueden proceder de dos o más variables y, entonces, se usa la expresión bivariante (si se trata de dos variables) o multivariante (si se consideran más de dos).

Gráficos y tablas para datos cualitativos o categóricos

Si se dispone de datos cualitativos o categóricos, pueden sintetizarse mediante una tabla que recoja, para cada categoría: el número de veces que aparece (frecuencia absoluta), el porcentaje de apariciones sobre el total de observaciones (frecuencia relativa), así como los acumulados de ambos valores. La tabla 1 muestra esta información para la variable "número de *hotspots* (conexiones *wi-fi*) identificados en cada comunidad autónoma".

Tabla 1. Ejemplo de tabla de frecuencias para una variable categórica

Comunidad autónoma	***Hotspots* por comunidad autónoma**			
	Frecuencia	**Frecuencia acumulada**	**Frecuencia relativa**	**Frec. rel. acumulada**
Andalucía	885	885	11,9%	11,9%
Aragón	177	1.062	2,4%	14,2%
Asturias	148	1.210	2,0%	16,2%
Cantabria	164	1.374	2,2%	18,4%
Castilla-La Mancha	144	1.518	1,9%	20,3%
Castilla y León	302	1.820	4,0%	24,4%
Cataluña	1.391	3.211	18,6%	43,0%
C. Valenciana	622	3.833	8,3%	51,3%
Extremadura	137	3.970	1,8%	53,2%
Galicia	516	4.486	6,9%	60,1%
I. Baleares	183	4.669	2,5%	62,5%
I. Canarias	151	4.820	2,0%	64,6%
La Rioja	126	4.946	1,7%	66,3%
Madrid	1.776	6.722	23,8%	90,0%
Murcia	160	6.882	2,1%	92,2%
Navarra	153	7.035	2,0%	94,2%
País Vasco	430	7.465	5,8%	100,0%
Totales	**7.465**		**100,0%**	

Nota

Observad que la **frecuencia acumulada** se obtiene sólo con ir acumulando frecuencias anteriores.

Además de mediante una tabla de frecuencias, suele ser habitual representar datos categóricos mediante el uso de gráficos circulares (figura 3) o bien mediante diagramas de barras (figura 4).

Figura 3. Ejemplo de gráfico circular para una variable categórica

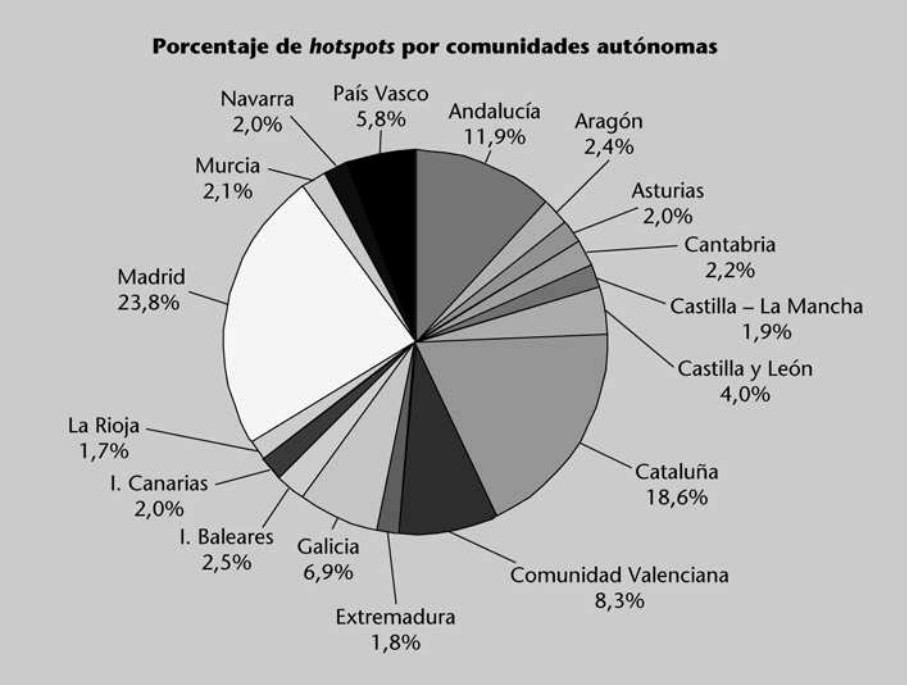
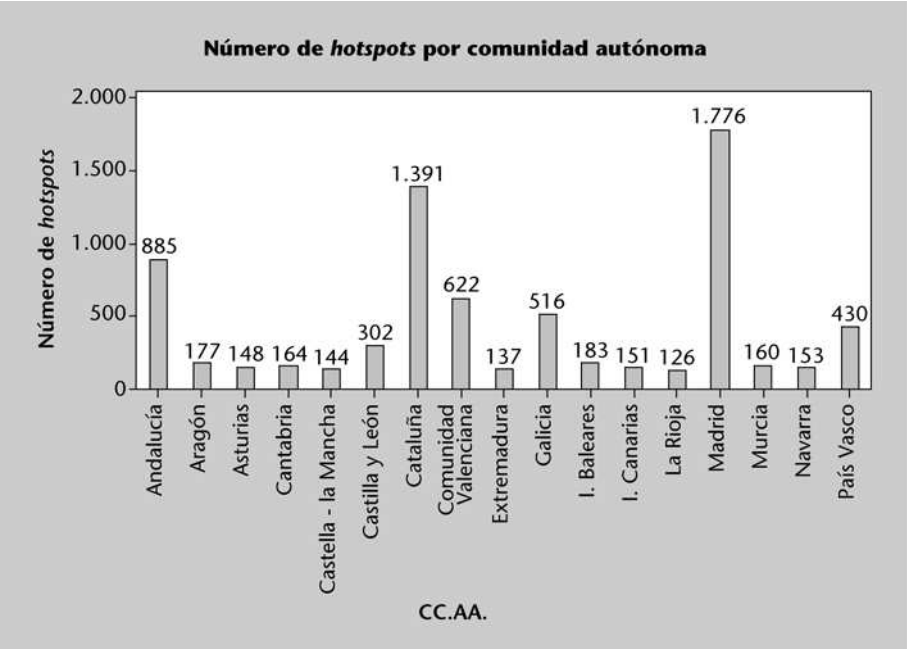
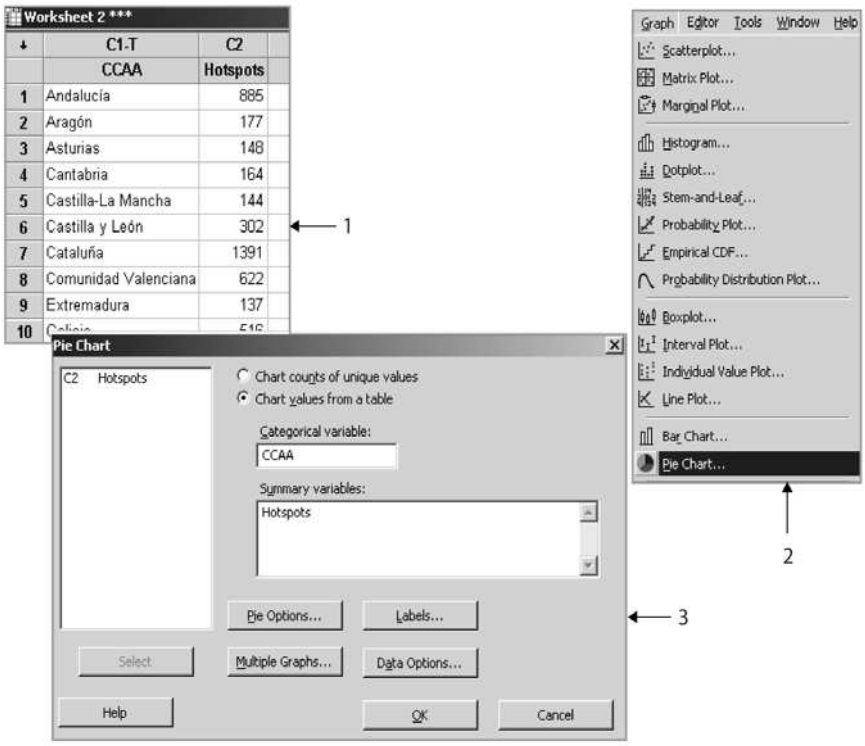


Figura 4. Ejemplo de diagrama de barras para una variable categórica



Este tipo de gráficos pueden crearse fácilmente con cualquier programa estadístico o de análisis de datos (p. ej.: Minitab, MS Excel, SPSS, etc.). La figura 5 muestra los pasos básicos para generar un gráfico circular (*pie chart*) con Minitab. La generación de un diagrama de barras (*bar chart*) se consigue de forma similar, al igual que ocurre con la mayoría de los gráficos que se presentan en este apartado.

Figura 5. Pasos a seguir para la generación de un gráfico circular con Minitab

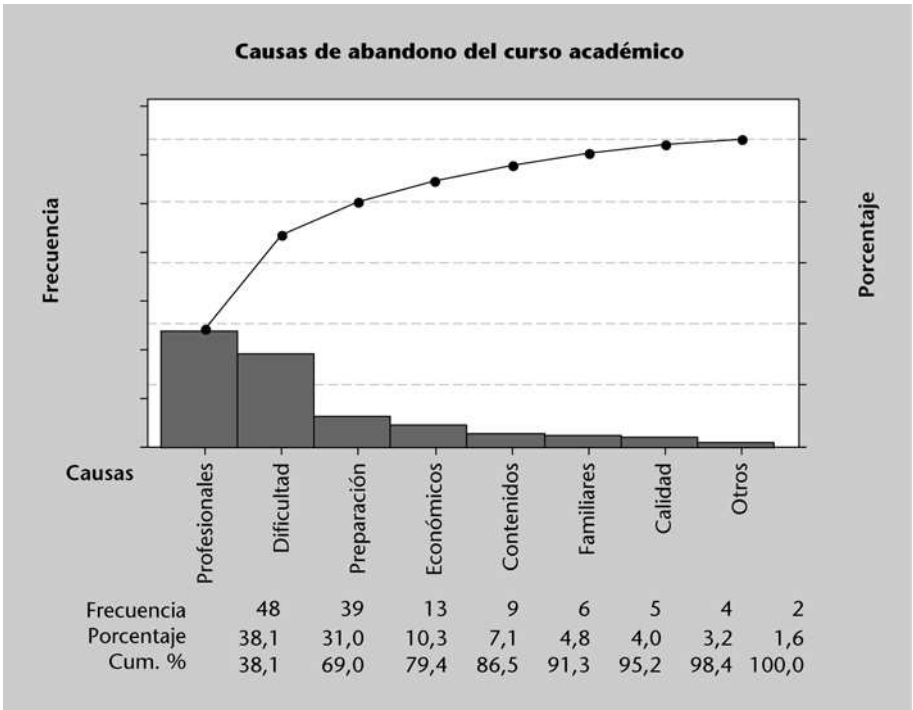


Un gráfico que también suele usarse bastante para describir datos cualitativos es el llamado diagrama de Pareto. Este gráfico está compuesto por: (a) un diagrama de barras en el que las categorías están ordenadas de mayor a menor frecuencia y (b) una línea que representa la frecuencia relativa acumulada (figura 6).

Pasos a seguir

Una vez introducidos los datos en el programa ***(1)***, se sigue la ruta ***Graph*** > ***Pie Chart (2)*** y se seleccionan las variables en la ventana correspondiente ***(3)***.

Figura 6. Diagrama de Pareto sobre las causas de abandono de un curso



Nota

Las capturas de pantalla de Minitab corresponden a la **versión 15** de este programa. Es posible que otras versiones ofrezcan ligeras diferencias en los menús y ventanas, aunque básicamente el proceso será el mismo. Para obtener más detalles sobre las opciones disponibles, siempre es posible consultar la ayuda en línea del programa o bien alguno de los numerosos manuales de uso que se pueden encontrar en Internet.

Diagrama de Pareto

Para generar un diagrama de Pareto en Minitab hay que usar la ruta *Stat* > *Quality Tools*.

Los diagramas de Pareto son muy útiles para detectar cuándo un porcentaje reducido de categorías (p. ej.: un 20% de las categorías) "acapara" o representa un porcentaje alto de observaciones (p. ej.: un 80% de los datos). Estos fenómenos de excesiva representatividad por parte de unas pocas categorías suelen darse con frecuencia en contextos socioeconómicos (p. ej.: un porcentaje reducido de los ciudadanos de un país acapara un alto porcentaje de la renta), educativos (p. ej.: un porcentaje reducido de causas generan la mayor parte de los abandonos del curso) o de ingeniería de la calidad (p. ej.: un alto porcentaje de fallos son debidos a un número muy reducido de causas). Identificar aquellas pocas categorías que representan una gran parte del porcentaje total puede servir para corroborar ciertos desequilibrios distributivos –como una distribución poco equilibrada de las rentas en un país o de los sueldos en una empresa–, o para proporcionar pistas sobre los principales factores de causa de un problema –como el alto nivel de abandono de un curso o un elevado nivel de fallos en un servicio o producto–.

Gráficos y tablas para datos cuantitativos

En el caso de datos cuantitativos, su representación gráfica o mediante tablas permite apreciar la forma de su distribución estadística, es decir, la forma en que se comporta la variable de interés (cuáles son los valores medios o centrales, cuáles son los valores más habituales, cómo varía, cómo de dispersos son los valores, si muestra algún patrón de comportamiento especial, etc.).

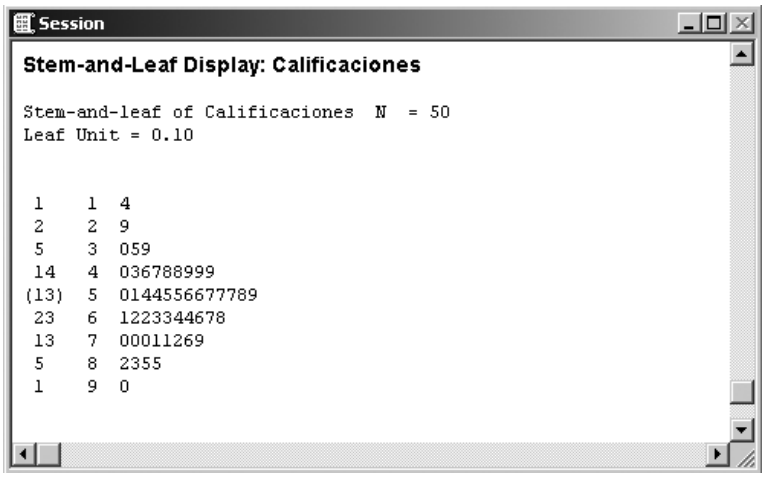
Uno de los gráficos más sencillos de elaborar es el llamado gráfico de puntos (*dotplot*). Se trata de un gráfico en el que cada punto representa una o más observaciones. Los puntos se apilan uno sobre otro cuando se repiten los valores observados (figura 7).

Figura 7. Gráfico de puntos para las calificaciones de un curso

Un gráfico similar, aunque algo más elaborado y con una orientación transpuesta de los ejes, es el llamado diagrama de tallos y hojas (*stem-and-leaf*). En él también se representan los valores observados pero usando los propios valores numéricos en lugar de puntos, lo que proporciona un mayor nivel de detalle. La figura 8 muestra un ejemplo de gráfico de tallos y hojas para los mismos datos empleados en la figura 7. Se observa que el gráfico se ha construido a partir de una muestra de cincuenta calificaciones y que

se ha usado una unidad de hoja (*leaf*) de 0,1. Esto significa que la segunda columna del gráfico representa la parte entera de la calificación, mientras que cada uno de los números situados a su derecha representa la parte decimal de una observación con dicha parte entera. Así, se pueden leer las siguientes calificaciones por orden de menor a mayor: 1,4, 2,9, 3,0, 3,5, 3,9, 4,0, 4,3, etc.

Figura 8. Gráfico de hojas y tallos para las calificaciones de un curso



Atención

Cabe destacar que en un gráfico de tallos y hojas los datos se apilan de izquierda a derecha en lugar de arriba abajo como ocurre con el gráfico de puntos.

Cuando las observaciones generan un número elevado de valores distintos, resulta recomendable agruparlos en clases o intervalos disjuntos de igual tamaño. De ese modo, cada observación se clasifica en una clase o intervalo según su valor. La tabla 2 muestra un ejemplo de tabla de frecuencias en el que se han agrupado los datos en intervalos. La frecuencia de cada intervalo viene determinada por el número de observaciones cuyos valores están en dicho intervalo. La marca de clase representa el valor medio del intervalo.

Tabla 2. Ejemplo de tabla de frecuencias agrupadas usando intervalos

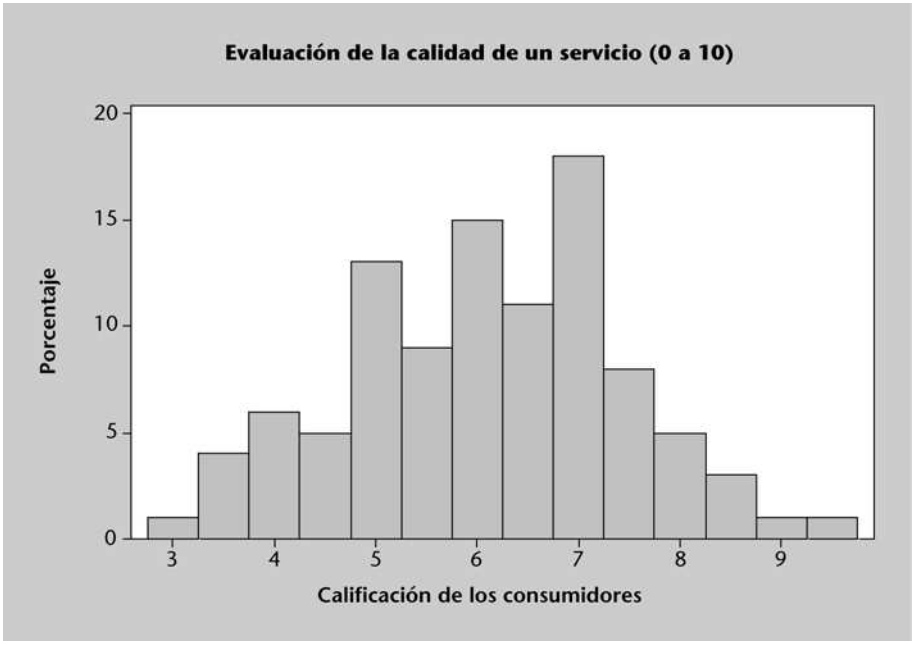
Intervalo	Marca de clase	Frecuencia	Frecuencia relativa
[0, 2)	1	12	8,1%
[2, 4)	3	23	15,5%
[4, 6)	5	67	45,3%
[6, 8)	7	31	20,9%
[8, 10)	9	15	10,1%
Totales		**148**	**100,0%**

Un gráfico que utiliza también intervalos para agrupar los datos a representar es el histograma. El histograma muestra la frecuencia (absoluta o relativa) de cada clase, lo que permite visualizar de forma aproximada la distribución de los datos (figura 9). Sin embargo, hay que tener presente que la forma final del histograma puede variar bastante según el número de intervalos que se definan para agrupar los datos, lo que a veces no permite apreciar correctamente la forma exacta de la distribución estadística que siguen las observaciones.

Nota

Una regla habitual es definir $\sqrt{n}$ clases o intervalos, siendo n el número de observaciones disponibles.

Figura 9. Histograma de una distribución aproximadamente normal



La figura 9 muestra un histograma con forma de campana: es una forma bastante simétrica, que presenta una mayor altura en la parte central y disminuye paulatinamente en las "colas" o extremos. Esta forma es bastante habitual y suele caracterizar el comportamiento de muchas variables (p. ej.: notas numéricas en un examen, peso o altura de individuos, temperaturas diarias, etc.). Sin embargo, también es habitual encontrarse con variables que muestran patrones de comportamientos completamente distintos. Por ejemplo, la figura 10 muestra un histograma en el que se aprecia una distribución más "uniforme" u homogénea de los datos, mientras que la figura 11 muestra un histograma en el que se aprecia una distribución asimétrica o "sesgada" de los mismos.

Figura 10. Histograma de una distribución aproximadamente uniforme

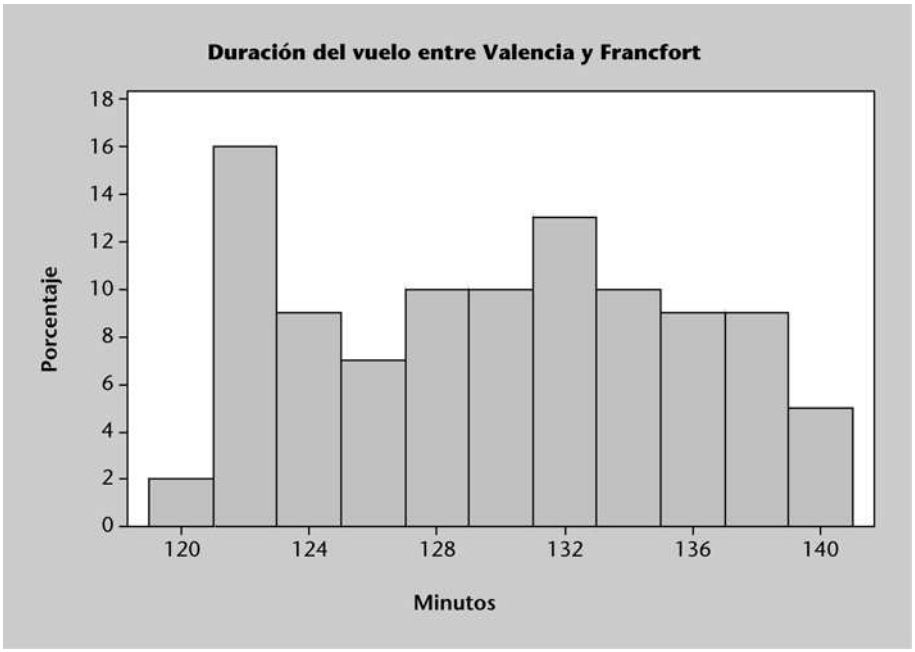
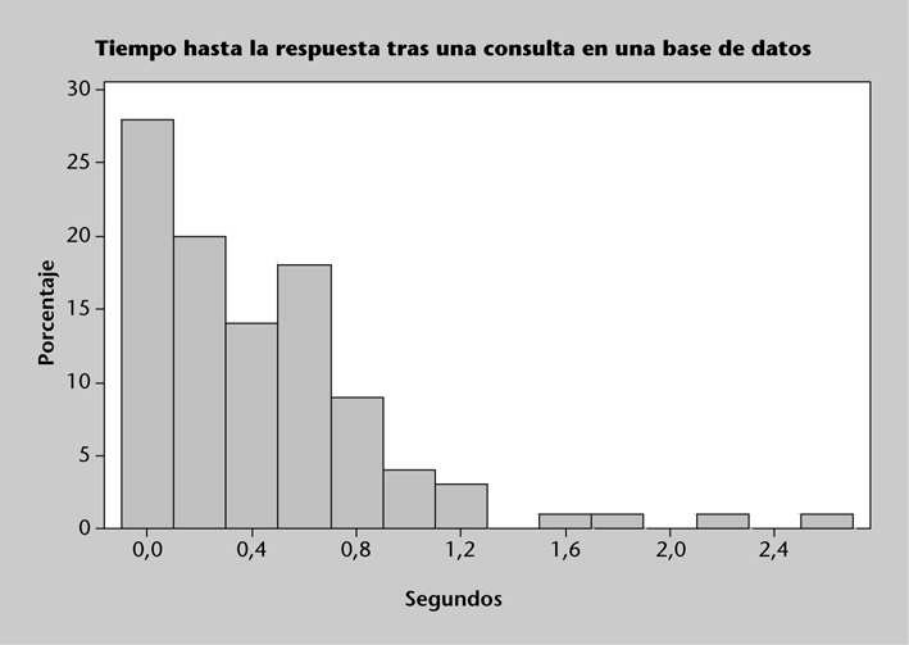


Figura 11. Histograma de una distribución sesgada a la derecha



3. Descripción de datos mediante estadísticos

Dado un conjunto de n datos u observaciones, x_1, x_2, …, x_n, asociadas a una variable de interés X, suele ser útil sintetizar algunas de sus principales propiedades en unos pocos valores numéricos. Los estadísticos descriptivos son, precisamente, estos valores numéricos capaces de proporcionar información a partir del conjunto de las observaciones. Estos estadísticos resultan muy útiles a la hora de entender el comportamiento de los datos, ya que un simple valor numérico es capaz de describir propiedades tan relevantes como, por ejemplo, el valor promedio del conjunto de datos, el valor máximo, el valor mínimo, el valor que se repite con más frecuencia, un índice de dispersión o variabilidad, etc.

Web

Recordar que la World Wide Web (p. ej., Wikipedia, etc.) es una excelente fuente de consulta para ampliar los conceptos y definiciones estadísticas que se proporcionan en este y otros módulos. Un recurso especialmente interesante, por cuanto ofrece una visión muy completa de conceptos y técnicas estadísticas, es el libro en línea de StatSoft http://www.statsoft.com/textbook/.

Como ya se comentó anteriormente, estos estadísticos hacen referencia a una muestra de observaciones y suelen representarse mediante letras del alfabeto latino ($\bar{x}$, s, etc.), lo que permite distinguirlos claramente de sus parámetros asociados que sintetizan propiedades de toda la población y se representan mediante letras griegas (μ, σ, etc.). Básicamente pueden distinguirse dos grupos de estadísticos descriptivos: (a) los de centralización, que proporcionan información sobre cuáles son los valores "centrales" del conjunto de datos (p. ej.: el valor promedio de los datos) y (b) los de dispersión, que explican cómo se sitúan y varían los datos con respecto a los valores "centrales" (p. ej.: el rango o diferencia entre el valor máximo y el valor mínimo de los datos).

Nota

Recordar que los símbolos μ y σ se pronuncian como "mu" y "sigma", respectivamente. La pronunciación de otros símbolos del alfabeto griego se puede consultar, p. ej., en Wikepedia.

Estadísticos de centralización

A continuación se presentan los estadísticos de centralización más usados habitualmente:

- **Media (*mean*)**: la media (también conocida por valor promedio o valor esperado) de un conjunto de observaciones muestrales se representa con el símbolo $\bar{x}$. Intuitivamente, la media simboliza el "centro de masas" o "punto de equilibrio central" del conjunto de datos considerado. El parámetro asociado, la media poblacional, se representa por μ. Para calcular la media de un conjunto de datos se usa la siguiente expresión:

Media muestral

Recordar que la media muestral es un ***estadístico*** que hace referencia al "centro de masas" de los datos de una muestra (subconjunto de la población), mientras que la media poblacional es un **parámetro** que representa el "centro de masas" de toda la población.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Ejemplo: la media de los cinco datos siguientes {6, 3, 8, 6, 4} es

$$\bar{x} = \frac{6+3+8+6+4}{5} = \frac{27}{5} = 5,4$$

- **Mediana (*median*)**: la mediana de un conjunto de observaciones muestrales suele representarse con el símbolo $\tilde{x}$. En el caso de una población, el

parámetro mediana se denota con *M*. Una vez se ordenan todos los datos de menor a mayor, la mediana es aquel valor que deja a su izquierda la mitad de las observaciones (es decir, es aquel valor tal que el número de observaciones más pequeñas que él coincide con el número de observaciones mayores que él). Los pasos para calcular la mediana son: (1) ordenar los datos de menor a mayor, (2) calcular la posición *i* que ocupa la mediana en el conjunto ordenado de datos, $i = \frac{n+1}{2}$ y (3) seleccionar la observación x_i (la que ocupa la posición determinada en el paso anterior). Cabe observar que si el número de datos *n* es impar (p. ej.: $n = 5$), la posición *i* será un valor entero (p. ej.: $i = 3$) que corresponderá con un valor concreto, x_i, del conjunto de datos. Sin embargo, si *n* es par (p. ej.: $n = 6$), la posición *i* será un número no entero (p. ej.: $i = 3{,}5$), en cuyo caso la mediana vendrá dada por el promedio de los dos valores que ocupan las posiciones enteras más cercanas a *i* (en este caso por el promedio de los valores que ocupan las posiciones 3 y 4).

Ejemplo: dado el conjunto de ocho datos {5, 11, 7, 8, 10, 9, 6, 9}, lo primero es ordenarlos de menor a mayor, con lo que se obtiene la serie {5, 6, 7, 8, 9, 9, 10, 11}; ahora, la posición de la mediana vendrá dada por $i = \frac{8+1}{2} = 4{,}5$, es decir, la mediana estará entre los valores que ocupan las posiciones 4 y 5, por lo que se calcula el promedio de ambos para dar el valor de la mediana, es decir: $\tilde{x} = \frac{8+9}{2} = 8{,}5$.

Es importante destacar que la media es muy sensible a la existencia de valores extremos (*outliers*), es decir, la inclusión o no de un valor que esté muy alejado del resto de los datos puede cambiar considerablemente el valor resultante de la media. Por el contrario, la mediana se ve mucho menos afectada por la presencia de dichos valores, lo que significa que la mediana es un "centro" más estable que la media en el sentido de que se ve menos afectado por la presencia de valores extremos en los datos.

- **Moda (*mode*):** la moda de un conjunto de datos es el valor que más veces se repite (el de mayor frecuencia).

  Ejemplo: la moda de la serie de datos {6, 3, 4, 8, 9, 6, 6, 3, 4} es 6, puesto que es el valor que más veces aparece en la serie.

Estadísticos de dispersión

Se presentan ahora los principales estadísticos de dispersión que, como se ha comentado anteriormente, proporcionan información sobre la variabilidad del conjunto de datos:

- **Rango (*range*):** el rango de un conjunto de datos es la diferencia entre el valor máximo y el mínimo de los mismos.

  Ejemplo: dado el conjunto de datos {2, 3, 8, 3, 5, 1, –8}, su rango es 8 – (–8) = 16

- **Varianza muestral (*sample variance*):** la varianza de una muestra se representa por el símbolo s^2. En el caso de una población, el parámetro varianza se representa con el símbolo σ^2. La varianza muestral será mayor cuanto mayor sean las diferencias entre cada una de las observaciones x_i y la media de los datos $\overline{x}$, en concreto:

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

  Esto significa que la varianza es una medida de la dispersión de los datos con respecto a su media, es decir, cuando menor sea la varianza, tanto más agrupados estarán los datos alrededor de su valor promedio. Por el contrario, cuanto mayor sea la varianza, tanto más dispersos estarán los datos.

  Ejemplo: la varianza muestral de la serie de 5 datos {6, 3, 8, 5, 3} es:

$$s^2 = \frac{(6-5)^2 + (3-5)^2 + (8-5)^2 + (5-5)^2 + (3-5)^2}{5-1} = 4,5$$

- **Desviación estándar (*standard deviation*):** la desviación estándar (o típica) de una muestra se representa con el símbolo *s*, mientras que la desviación estándar de una población se representa con σ. La desviación estándar es la raíz cuadrada positiva de la varianza, esto es: $s = \sqrt{s^2}$ (o, dicho de otro modo, la varianza es el cuadrado de la desviación estándar).

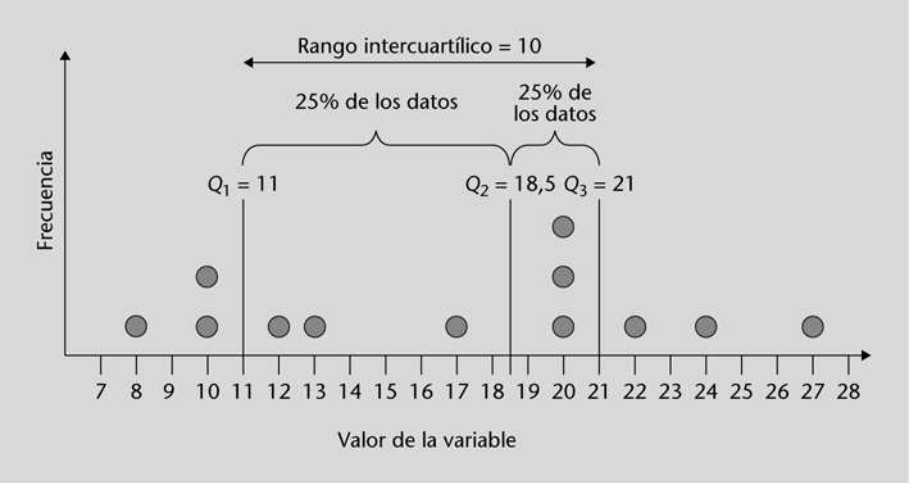
  Ejemplo: para los datos del ejemplo anterior, $s = \sqrt{4,5} = 2,1$

  Al igual que ocurría con la varianza, a mayor desviación estándar más dispersión en los datos y viceversa.

- **Cuartiles (*quartiles*):** en un conjunto de *n* observaciones ordenadas de menor a mayor valor, se pueden considerar tres valores numéricos concretos llamados cuartiles que dividen el conjunto en cuatro partes, cada una de ellas conteniendo una cuarta parte de las observaciones (figura 12). El primer cuartil, Q_1, es el valor que deja la cuarta parte de los datos ordenados a su izquierda (es decir, un 25% de los datos muestran valores inferiores a él y un 75% de los datos muestran valores superiores a él). Por su parte, el segundo cuartil, Q_2, es aquel valor que deja la mitad de los datos ordenados a su izquierda (es decir, un 50% de los datos muestran valores inferiores a él y un 50% de los datos muestran valores superiores a él). Finalmente, el tercer cuartil, Q_3, es aquel va-

lor que deja tres cuartas partes de los datos ordenados a su izquierda (es decir, un 75% de los datos muestran valores inferiores a él y un 25% de los datos muestran valores superiores a él).

Figura 12. Cuartiles de un conjunto ordenado de datos



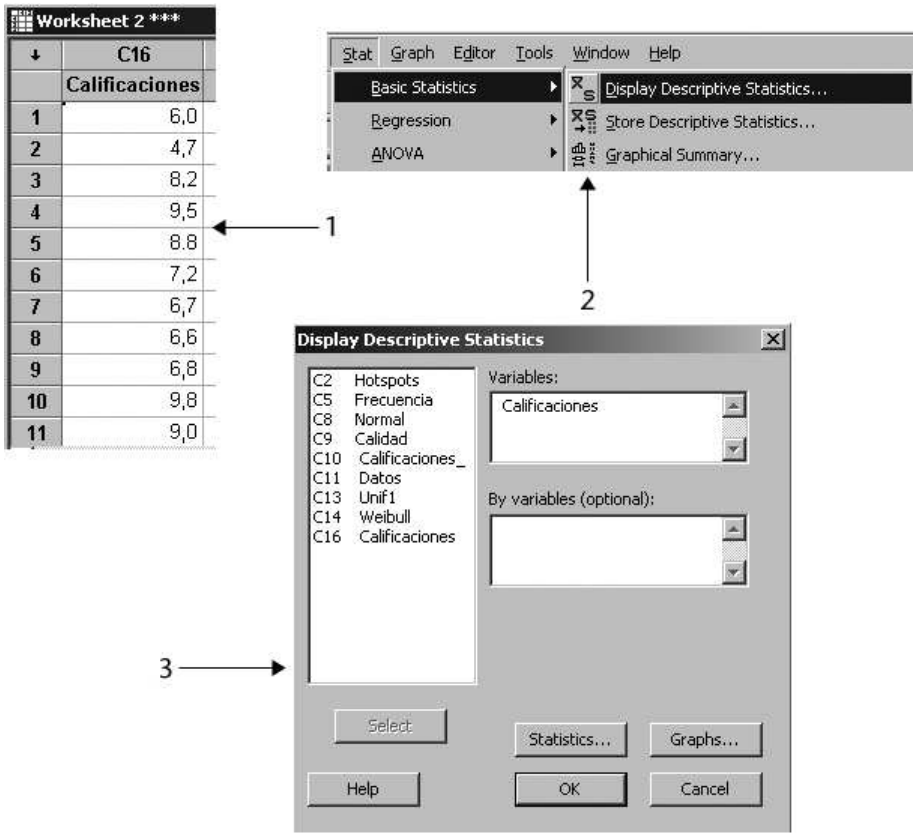
Obsérvese que, en realidad, el cuartil segundo o $\boldsymbol{Q_2}$ coincide con el concepto de mediana presentado anteriormente. Los cuartiles son muy útiles a la hora de clasificar una observación en una determinada franja del conjunto de datos, por ejemplo, si la observación es inferior a $\boldsymbol{Q_1}$ significa que ésta se encuentra situada entre el 25% de valores más bajos; si la observación es superior a $\boldsymbol{Q_3}$ significa que está situada entre el 25% de valores más altos, etc.

- **Rango intercuartílico (*inter-quartilic range*):** este rango suele representarse como *IQR* y es simplemente la diferencia entre el tercer cuartil y el primer cuartil, es decir: $\boldsymbol{IQR = Q_3 - Q_1}$. El rango intercuartílico indica el espacio que ocupan el 50% de las observaciones "centrales" (figura 12), por lo que, de forma similar a lo que ocurría con la varianza, da una medida de la dispersión de los datos (a mayor *IQR* mayor dispersión y viceversa).

Obtención de estadísticos descriptivos mediante programas informáticos

En la práctica, es habitual utilizar algún programa estadístico o de análisis de datos para calcular los estadísticos anteriores e incluso algunos estadísticos adicionales que proporcionen información sobre el conjunto de datos. En la figura 13 se muestran los pasos básicos necesarios para obtener los principales estadísticos descriptivos con Minitab. El *output* del programa, para un ejemplo con cincuenta observaciones, se muestra en la figura 14. Por su parte, la figura 15 muestra una serie de estadísticos descriptivos generados con MS Excel para el mismo conjunto de datos (en este caso los cuartiles se han obtenido usando las fórmulas integradas de Excel).

Figura 13. Pasos para calcular estadísticos descriptivos con Minitab



Pasos a seguir

Una vez introducidos los datos en el programa *(1)*, se sigue la ruta ***Stat*** > ***Basic Statistics*** > ***Display Descriptive Statistics*** *(2)* y se seleccionan las variables en la ventana correspondiente *(3)*.

Figura 14. Estadísticos descriptivos obtenidos con Minitab

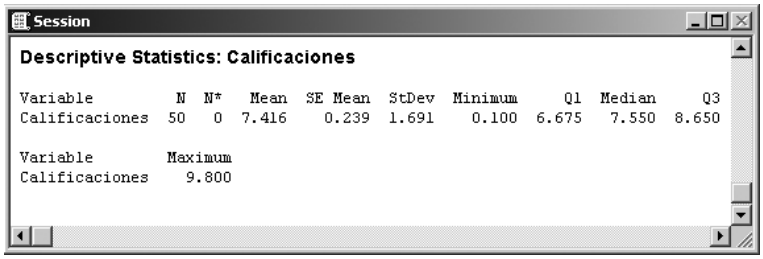


Figura 15. Estadísticos descriptivos calculados con Excel

	A	B	C	D
1	Calificaciones			
2	6,0		*Calificaciones*	
3	4,7			
4	8,2		Media	7,416
5	9,5		Error típico	0,239205488
6	8,8		Mediana	7,55
7	7,2		Moda	7,2
8	6,7		Desviación estándar	1,691438224
9	6,6		Varianza de la muestra	2,860963265
10	6,8		Curtosis	5,922785035
11	9,8		Coeficiente de asimetría	-1,720324942
12	9,0		Rango	9,7
13	7,7		Mínimo	0,1
14	8,6		Máximo	9,8
15	5,8		Suma	370,8
16	6,4		Cuenta	50
17	9,5			
18	7,4		Cuartil primero	6,725
19	7,2		Cuartil segundo	7,55
20	8,8		Cuartil tercero	8,6
21	7,4			

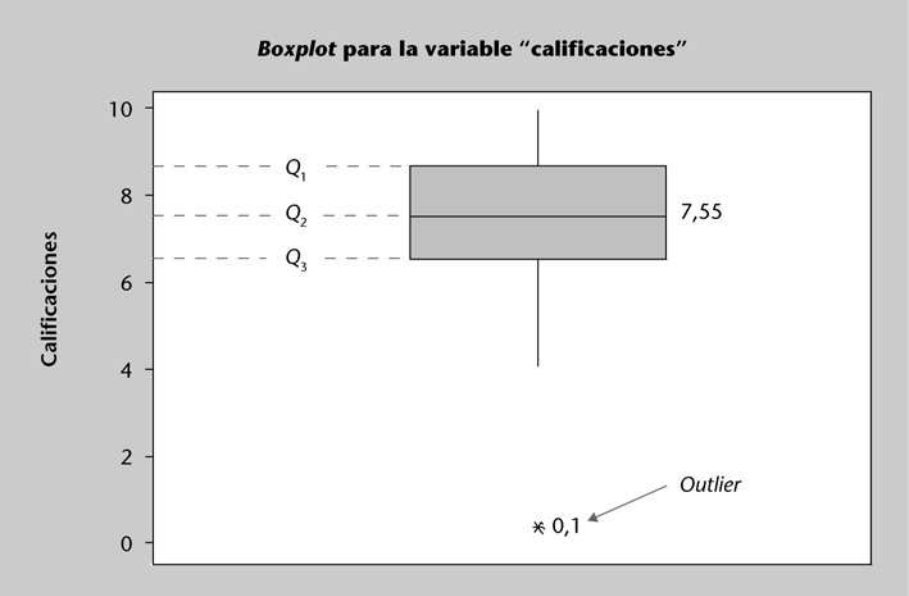
Diferencias en los métodos de cálculos

Cabe destacar que hay ligeras diferencias entre los valores de los cuartiles calculados por Minitab y los correspondientes valores de Excel. Ello se debe a que usan métodos de cálculo distintos. Una discusión interesante sobre los diferentes métodos existentes para calcular los cuartiles se puede encontrar en: http://mathforum.org/library/drmath/view/60969.html.

Diagrama de cajas y bigotes (*boxplot*)

Usando los cuartiles es posible construir un tipo de gráfico, el diagrama de cajas y bigotes (*boxplot*), que resulta muy útil para visualizar la distribución de los datos. Este diagrama está compuesto por una caja central, definida por los cuartiles primero y tercero, que contiene el 50% "central" de las observaciones, y dos segmentos situados en los respectivos extremos de la caja, representando cada uno de ellos el 25% de las observaciones extremas (figura 16).

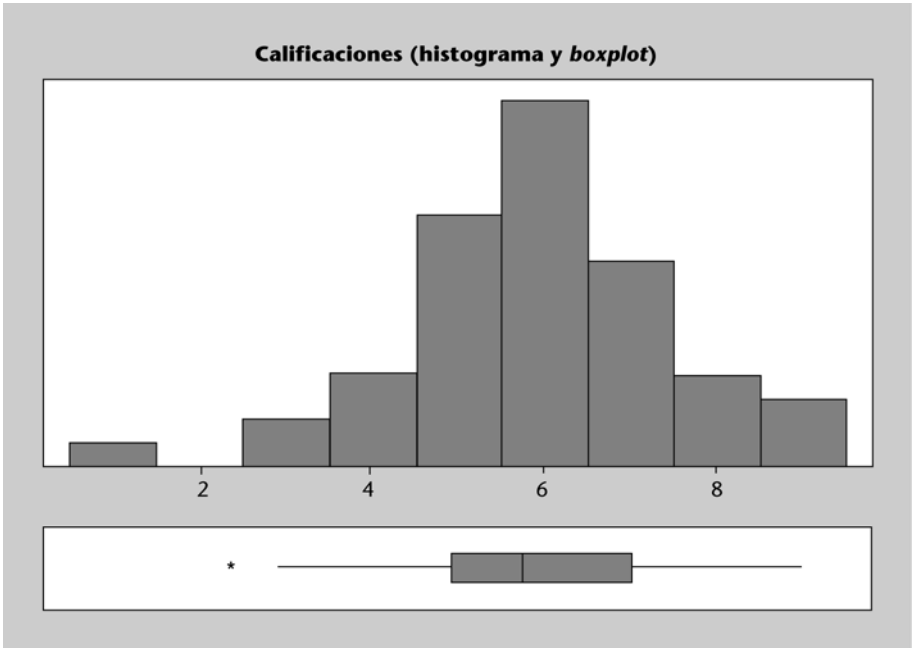
Figura 16. Diagrama de cajas y bigotes (*boxplot*) y valores extremos (*outliers*)



El diagrama de cajas y bigotes sirve también para identificar posibles valores anómalos (*outliers*), que se encuentran excesivamente alejados del resto de los datos, es decir: o bien son extremadamente grandes o bien extremadamente pequeños en comparación con el resto de observaciones. Estos valores anómalos se suelen representar mediante un asterisco, y pueden ser debidos a un error en el registro de los datos o bien a valores que, en realidad, se encuentran extremadamente alejados del resto de observaciones (p. ej.: el precio de un Ferrari cuando se compara con precios de turismos de gama media). Identificar valores anómalos en un conjunto de observaciones es importante, puesto que el análisis de los datos puede dar resultados muy distintos en función de que se consideren o no dichos valores en el estudio (por ejemplo, la media y la varianza de un conjunto de datos pueden cambiar de forma notable según se incluya o no uno de estos valores extremos).

La estrecha relación existente entre el histograma y el *boxplot* se puede observar en la figura 17. En cierto sentido, el *boxplot* se puede interpretar como un histograma visto desde arriba. En este caso, la zona del *boxplot* situada entre los cuartiles primero y tercero correspondería a la zona central del histograma. Además, en ambos casos queda identificado el valor anómalo (*outlier*) así como la forma aproximadamente simétrica del resto de la distribución.

Figura 17. Relación entre histograma y *boxplot*



4. El concepto de probabilidad

Un **experimento aleatorio** es aquel en el que no es posible conocer a priori el suceso resultante que acontecerá pero, sin embargo, sí es posible observar un cierto patrón regular en los resultados que van sucediendo cuando el experimento se repite muchas veces. Por ejemplo, cuando se considera el experimento aleatorio consistente en lanzar una moneda (o un dado) al aire, no es posible predecir cuál será el **suceso resultante** del experimento, es decir, si saldrá cara o cruz (o qué número saldrá en el caso del dado); sin embargo, sí se puede afirmar que tras muchos lanzamientos el porcentaje o proporción de sucesos "cara" obtenidos será muy próximo al 50% o 1/2 (en el caso del dado, el porcentaje o proporción de sucesos "3" obtenidos será muy próximo a 0,1667 o 1/6). Este porcentaje o proporción de aparición de un suceso tras muchas repeticiones del experimento es lo que da lugar a la idea de probabilidad:

Ejemplo

La **probabilidad** de un suceso es siempre un número entre 0 y 1. Así, por ejemplo, una probabilidad de 0,25 representa un porcentaje de aparición del 25% o, equivalentemente, una proporción de 1/4.

> Se define la **probabilidad de un suceso** A, $P(A)$, como el porcentaje o proporción de aparición de dicho suceso en una serie extraordinariamente larga de repeticiones del experimento, todas ellas independientes entre sí.

El requisito de independencia entre las distintas repeticiones del experimento aleatorio significa que el resultado de cada repetición del experimento no está condicionado por los resultados obtenidos en repeticiones anteriores (p. ej.: cuando se lanza varias veces una moneda al aire, el suceso resultante de cada nuevo lanzamiento es independiente de los resultados obtenidos en lanzamientos previos).

Ejemplo 1 de probabilidades

En el experimento "lanzamiento de una moneda al aire", es posible considerar los siguientes sucesos o potenciales resultados: C = {cara}, X = {cruz}, Ω = {cara o cruz} y $\varnothing$ = {ni cara ni cruz}. Los dos últimos sucesos se conocen, respectivamente, como suceso seguro Ω (que incluye todos los resultados posibles) y suceso imposible o conjunto vacío $\varnothing$ (que no incluye ningún resultado derivado de la ejecución del experimento). En este caso, parece claro que $P(C) = 0{,}5$ (es decir, si se repitiera el experimento muchas veces, aproximadamente el 50% de las mismas serían caras), $P(X) = 0{,}5$, $P(\Omega) = 1$ (es decir, en el 100% de los lanzamientos saldrá o bien cara o bien cruz) y $P(\varnothing) = 0$ (es decir, en el 0% de los lanzamientos no se obtendrá resultado alguno).

Ejemplo 2 de probabilidades

En el experimento aleatorio "lanzamiento de un dado", es posible considerar sucesos o potenciales resultados como los siguientes: {1}, {2}, {3}, {4}, {5}, {6},

Ω = {un número entre 1 y 6}, $\varnothing$ = {ningún número entre 1 y 6}. En este caso, $P(\{1\}) = 1/6$ (tras muchas repeticiones, uno de cada seis lanzamientos acabará siendo un 1), $P(\{2\}) = 1/6$, $P(\{3\}) = 1/6$, $P(\{4\}) = 1/6$, $P(\{5\}) = 1/6$, $P(\{6\}) = 1/6$, $P(\Omega) = 1$ y $P(\varnothing) = 0$.

Observar, además, que también es posible considerar sucesos compuestos como, por ejemplo, par = {2, 4, 6}, impar = {1, 3, 5}, mayor2 = {3, 4, 5, 6}, menor3 = {1, 2}, etc. En este caso, $P(\text{par}) = 3/6 = 1/2$, $P(\text{impar}) = 1/2$, $P(\text{mayor2}) = 4/6 = 2/3$, $P(\text{menor3}) = 2/6 = 1/3$.

Propiedades básicas de las probabilidades

Hay una serie de propiedades básicas que debe satisfacer cualquier probabilidad. Estas propiedades son muy útiles a la hora de calcular probabilidades de sucesos complejos a partir de probabilidades ya conocidas o fáciles de obtener:

1) La probabilidad de cualquier suceso A siempre es un número situado entre 0 y 1 (ambos inclusive), es decir $0 \leq P(A) \leq 1$.

Ejemplo: en los ejemplos anteriores, todas las probabilidades halladas eran valores entre 0 y 1.

2) La probabilidad del suceso imposible o conjunto vacío $\varnothing$ es siempre 0, es decir, $P(\varnothing) = 0$. En otras palabras, cuando se hace un experimento aleatorio siempre se obtiene algún resultado y, por tanto, la proporción de "no-resultados" es 0.

Ejemplo: en los ejemplos anteriores, $P(\varnothing) = 0$.

3) La suma de las probabilidades de todos los posibles resultados del experimento aleatorio siempre vale 1. En otras palabras, la probabilidad del suceso seguro es siempre 1.

Ejemplo: En el ejemplo de la moneda, $P(\Omega) = 1 = P(C) + P(X)$; en el ejemplo del dado, $P(\Omega) = 1 = P(\{1\}) + P(\{2\}) + P(\{3\}) + P(\{4\}) + P(\{5\}) + P(\{6\})$.

4) La probabilidad de que un suceso no ocurra es 1 menos la probabilidad de que sí ocurra, es decir: $P(\textit{no } A) = 1 - P(A)$.

Ejemplo: en el ejemplo de la moneda, $P(C) = 0{,}5 = 1 - P(\textit{no } C) = 1 - P(X)$; en el ejemplo del dado, $P(\text{par}) = 0{,}5 = 1 - P(\text{no par}) = 1 - P(\text{impar})$; $P(\varnothing) = 1 - P(\Omega)$.

5) Si dos sucesos A y B no tienen resultados comunes (son disjuntos), la probabilidad de que ocurra $A \cup B$ es la suma de las probabilidades, es decir, si A y B son disjuntos, $P(A \cup B) = P(A) + P(B)$.

Ejemplo: en el ejemplo de la moneda, $P(C \cup X) = P(C) + P(X) = 1$; en el ejemplo del dado, $P(\{1, 2\}) = P(\{1\}) + P(\{2\}) = 2/6 = 1/3$; $P(\Omega \cup \varnothing) = P(\Omega) + P(\varnothing) = 1 + 0 = 1$.

6) En general, para cualesquiera dos sucesos A y B se cumplirá que $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, donde "$A \cap B$" es el conjunto de posibles resultados que satisfacen los sucesos A y B a la vez. Hay que tener en cuenta que cuando A y B son disjuntos (no tienen resultados en común), "$A \cap B$" = $\varnothing$ y, por tanto, $P(A \cup B) = P(A) + P(B) - P(\varnothing) = P(A) + P(B) - 0 = P(A) + P(B)$, que es la expresión vista en la propiedad anterior.

Ejemplo: en el ejemplo del dado, $P(\text{par} \cup \text{mayor2}) = P(\text{par}) + P(\text{mayor2}) - P(\text{par} \cap \text{mayor2}) = 3/6 + 4/6 - 2/6 = 5/6$ (observar que "par $\cap$ mayor2" = {4, 6}).

5. Distribuciones de probabilidad discretas

Al inicio de este módulo se definió el concepto de variable cuantitativa discreta como aquella variable cuantitativa que podía tomar un número finito o contable de valores distintos. Así, un ejemplo de variable discreta sería X = "resultado del lanzamiento de un dado", ya que dicha variable sólo puede tomar seis posibles valores.

Cada uno de los posibles valores de una variable discreta tendrá asociada una probabilidad de ocurrencia (p. ej., en el caso del dado, la probabilidad de obtener un 2 será de 1/6), por lo que parece natural estudiar cómo se distribuyen o comportan dichas probabilidades. En concreto, se puede definir una "función de probabilidad", $f(x)$, que asocie a cada valor x de la variable discreta X su probabilidad de ocurrencia, $P(x)$. Por ejemplo, en el caso de la variable anterior, asociada al experimento aleatorio **"lanzamiento de un dado normal"**, la correspondiente función de probabilidad sería: $f(1) = P(X = 1) = 1/6$, $f(2) = P(X = 2) = 1/6$, $f(3) = P(X = 3) = 1/6$, $f(4) = P(X = 4) = 1/6$, $f(5) = P(X = 5) = 1/6$, $f(6) = P(X = 6) = 1/6$.

Observad

Fijaos que si se usara un **dado "trucado"**, no todas las probabilidades de ocurrencia serían iguales y, por tanto, la función de probabilidad tomaría valores distintos para distintos valores posibles de la variable.

Dada una variable aleatoria discreta X, resulta útil conocer la **distribución de probabilidad** de dicha variable, es decir, cómo se distribuyen o comportan las probabilidades de ocurrencia de sus posibles valores. A tal efecto se definen las siguientes funciones:

La **función de probabilidad** de X es aquella función $f(x)$ que asigna a cada posible valor x de X su probabilidad de ocurrencia, es decir: $f(x) = P(X = x)$ para todo valor posible x de X.

La **función de distribución** de X es aquella función $F(x)$ que asigna a cada posible valor x de X su probabilidad acumulada de ocurrencia, es decir $F(x) = P(X \leq x)$ para todo valor posible x de X.

La tabla 3 muestra la función de probabilidad y la función de distribución correspondientes a la variable X anterior pero usando un dado "trucado" que tiene dos valores 6 y ningún valor 2. Por su parte, la figura 18 muestra ambas funciones superpuestas en el mismo gráfico. Observando detenidamente la tabla 3 y la figura 18 se pueden deducir las siguientes características propias de estas funciones:

- Puesto que representan probabilidades, ambas funciones siempre toman valores en el intervalo [0, 1].

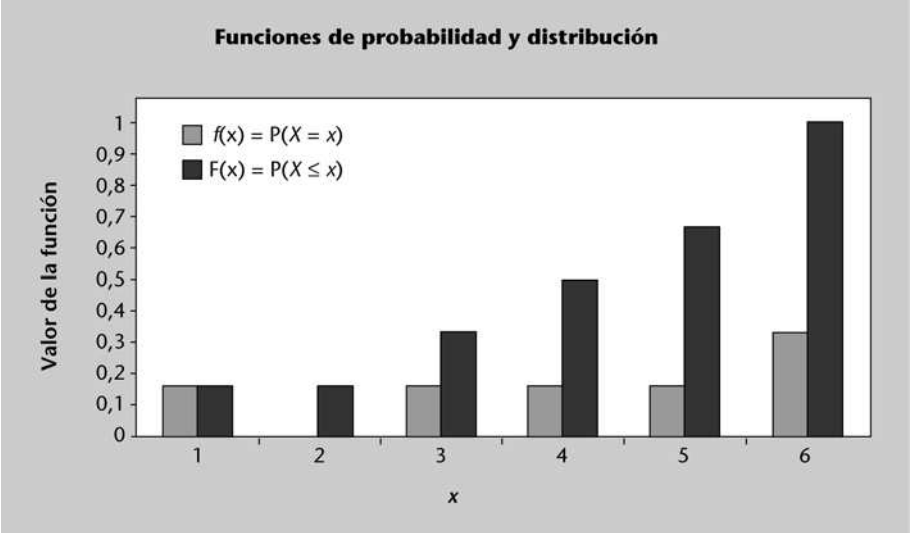
- La suma de todos los valores que toma la función de probabilidad siempre ha de ser 1 (ello se debe a las propiedades de la probabilidad).

La función de distribución siempre es una función creciente que pasa de valor 0 en su extremo izquierdo ($F(0) = P(X \leq 0) = 0$) a valor 1 en su extremo derecho ($F(6) = P(X \leq 6) = 1$).

Tabla 3. Funciones de probabilidad y distribución para una variable discreta

Variable X	Función de probabilidad $f(x) = P(X = x)$	Función de distribución $F(x) = P(X \leq x)$
1	1/6	1/6
2	0	1/6
3	1/6	2/6
4	1/6	3/6
5	1/6	4/6
6	2/6	1
Total	**1**	

Figura 18. Funciones de probabilidad y distribución de una variable discreta



Parámetros descriptivos de una distribución discreta

Mientras que los estadísticos descriptivos y los gráficos o tablas de frecuencias se utilizan para analizar el comportamiento (distribución) de una muestra de observaciones empíricas, las distribuciones de probabilidad son modelos estadísticos que usan parámetros y funciones de distribución para describir el comportamiento teórico (distribución teórica) de toda una población. De forma análoga a lo que ocurría con las muestras –que se caracterizan por estadísticos descriptivos como la media o la varianza muestral–, las distribuciones de probabilidad asociadas a poblaciones también suelen caracterizarse por parámetros tales como la media o la varianza poblacional. Ahora bien, puesto que en general no se dispondrá de observaciones sobre toda la población sino sólo de una función de distribución o de probabilidades, la forma de calcular dichos parámetros es algo distinta:

- **Media o valor esperado de una variable discreta:** la media o valor esperado de una variable discreta X que puede tomar los valores x_1, x_2, ..., se representa con μ o $E[X]$ y se calcula de la siguiente forma:

$$\mu = E[X] = x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + ... = \sum_i x_i \cdot f(x_i)$$

donde $f(x)$ denota a la función de probabilidad de X.

Ejemplo: el caso de un dado equilibrado, el valor esperado o media de X = "resultado del lanzamiento" sería $\mu = 3$; sin embargo, en el caso del dado "trucado" que se muestra en la tabla 3, la media o valor esperado es:

$$\mu = 1 \cdot f(1) + 2 \cdot f(2) + 3 \cdot f(3) + 4 \cdot f(4) + 5 \cdot f(5) + 6 \cdot f(6) =$$
$$= 1 \cdot \frac{1}{6} + 2 \cdot 0 + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{2}{6} = 4{,}167$$

- **Varianza y desviación estándar de una variable discreta:** la varianza de una variable discreta X que puede tomar los valores x_1, x_2, ..., se representa con σ^2 y se calcula de la siguiente forma:

$$\sigma^2 = (x_1 - \mu)^2 \cdot P(X = x_1) + (x_2 - \mu)^2 \cdot P(X = x_2) + ... = \sum_i (x_i - \mu)^2 \cdot f(x_i)$$

donde $f(x)$ denota a la función de probabilidad de X. De forma análoga a cómo ocurría con los estadísticos muestrales, la desviación estándar de una variable es la raíz cuadrada positiva de su varianza, es decir:

$$\sigma = \sqrt{\sigma^2}$$

Ejemplo: en el caso del dado "trucado" que se muestra en la tabla 3, la varianza es:

$$\sigma^2 = (1 - 4{,}167)^2 \cdot \frac{1}{6} + (2 - 4{,}167)^2 \cdot 0 + (3 - 4{,}167)^2 \cdot \frac{1}{6} +$$
$$+ (4 - 4{,}167)^2 \cdot \frac{1}{6} + (5 - 4{,}167)^2 \cdot \frac{1}{6} + (6 - 4{,}167)^2 \cdot \frac{2}{6} = 3{,}139$$

Y la correspondiente desviación estándar: $\sigma = \sqrt{3{,}139} = 1{,}772$

La distribución binomial

Una de las distribuciones discretas más usadas en la práctica es la distribución binomial. Esta distribución se usa para contestar a preguntas como las siguientes:

- Si cada vez que un sistema informático es atacado por un virus la probabilidad de que el sistema no falle es de 0,76, ¿cuál es la probabilidad de que no se haya producido ningún fallo en el sistema tras cinco ataques?

- Si cada vez que se consulta una fuente de información la probabilidad de que ésta proporcione una respuesta satisfactoria es de 0,85, ¿cuál es la probabilidad de que se obtenga alguna respuesta satisfactoria tras tres consultas?

- Si tras la administración de un fármaco a un paciente en estado crítico la probabilidad de supervivencia de éste es de 0,99, ¿cuál es la probabilidad de que sobrevivan los catorce pacientes críticos que han recibido el tratamiento?

- Si la probabilidad de obtener una concesión para un proyecto de investigación es de 0,20, ¿cuál es la probabilidad de obtener al menos una concesión tras tres intentos?

- Si cada vez que se trata de encuestar a un transeúnte elegido al azar la probabilidad de que responda es de 0,15, ¿cuál es la probabilidad de que se consigan obtener ochenta respuestas o más a partir de una muestra aleatoria de ciento cincuenta transeúntes?

Distribución de Poisson y la uniforme discreta

Otras distribuciones discretas muy habituales son la distribución de Poisson y la uniforme discreta. Es posible encontrar en Internet abundante documentación sobre éstas y otras distribuciones discretas así como sobre sus ámbitos de aplicación.

La **distribución binomial** es un modelo estadístico que permite calcular probabilidades sobre la variable aleatoria X = "número de éxitos conseguidos en n pruebas independientes". Cada una de estas n pruebas es una repetición de un experimento aleatorio cuyo resultado es binario (éxito o fracaso), siendo p la probabilidad de "éxito" en cada prueba y $q = 1 - p$ la probabilidad de "fracaso".

Resultado "éxito"

No debe confundirse el resultado "**éxito**" de un experimento aleatorio con el hecho de que el resultado sea deseable desde un punto de vista social o subjetivo. Así, por ejemplo, se podría considerar "éxito" del experimento aleatorio el fallo del sistema informático que sufre el ataque de un virus.

Cabe observar que la variable X = "número de éxitos en n pruebas independientes" puede tomar cualquier valor k entre 0 y n (ambos inclusive). Se suele usar la notación $X \sim B(n, p)$ para indicar que X se distribuye o se comporta según una distribución binomial de parámetros n (número de pruebas o repeticiones) y p (probabilidad de "éxito" en cada prueba). En tales condiciones, las probabilidades asociadas a dicha variable vienen dadas por la expresión matemática siguiente:

Para cualquier k entre 0 y n, $P(X = k) = \binom{n}{k} p^k \cdot (1-p)^{n-k}$, donde $\binom{n}{k} = \frac{n!}{k!(n-k)!}$,

siendo $0! = 1! = 1$ y $n! = n \cdot (n-1) \ldots 1$ para todo $n > 1$.

Observad

La expresión "$n!$" se lee como "factorial de n" o "n factorial". Así, por ejemplo, $4! = 4 \cdot 3 \cdot 2 \cdot 1$ y $6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$. Sin embargo, $1! = 1$ y $0! = 1$.

Se cumple, además, que la media (valor esperado) y la varianza de una distribución binomial son, respectivamente: $\mu = n \cdot p$ y $\sigma^2 = n \cdot p \cdot (1-p)$.

Ejemplo: la probabilidad de que al introducir datos en un formulario web se cometa un error es de 0,1. Si diez personas rellenan el formulario de forma independiente, ¿cuál es la probabilidad de que no haya más de un formulario erróneo?, ¿cuál es el valor esperado y la desviación estándar de la variable considerada?

Fijémonos en que, en este caso, X = "número de formularios erróneos en diez pruebas" y $X \sim B(10, 0{,}1)$. Además, se pide $P(X \leq 1) = P(X = 0 \cup X = 1) = P(X = 0) + P(X = 1)$ (puesto que son sucesos disjuntos). Ahora bien:

$$P(X=0)=\binom{10}{0}0{,}1^0\cdot(0{,}9)^{10}=\frac{10!}{0!10!}(1)(0{,}3487)=0{,}3487$$

$$P(X=1)=\binom{10}{1}0{,}1^1\cdot(0{,}9)^{9}=\frac{10!}{1!9!}(0{,}1)(0{,}3874)=0{,}3874$$

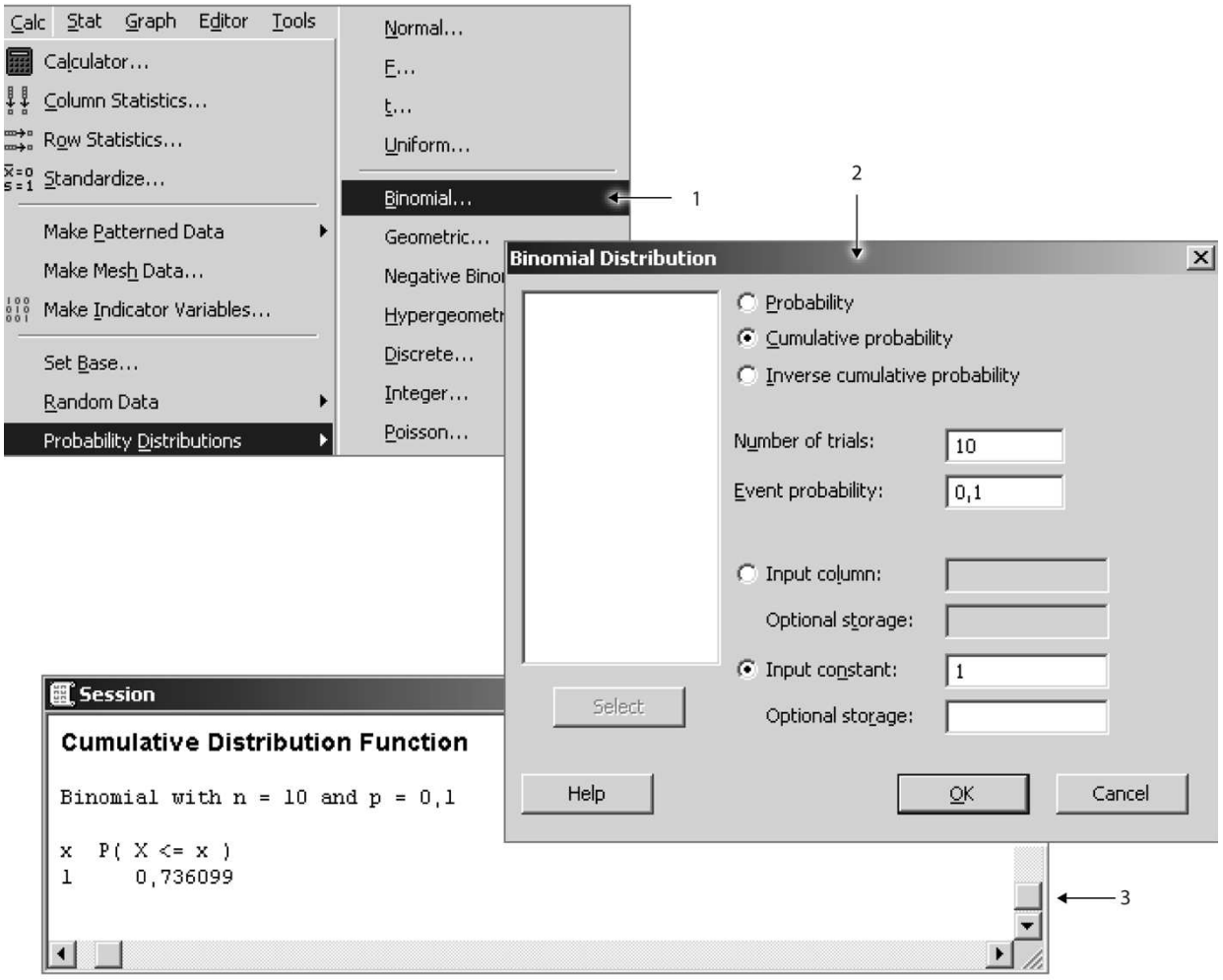
Por tanto, $P(X \leq 1) = 0{,}3874 + 0{,}3487 = 0{,}7361$. Finalmente, $\mu = 10 \cdot 0{,}1 = 1$ y $\sigma = \sqrt{10 \cdot 0{,}1 \cdot 0{,}9} = 0{,}9487$.

Pasos a seguir

Se sigue la ruta ***Calc*** > ***Probability Distributions*** > ***Binomial (1)*** y se completan los parámetros en la ventana correspondiente ***(2)***. El resultado se muestra en ***(3)***. Observar que, si en lugar de escoger la opción ***Cumulative probability*** en ***(2)*** se hubiera escogido la opción ***Probability***, el programa hubiera calculado $P(X = 1)$ en lugar de $P(X <= 1)$. Finalmente, para una probabilidad p dada, la opción ***Inverse cumulative probability*** devuelve aquel valor c de la variable X tal que $P(X <= c) = p$.

En la práctica, los cálculos probabilísticos anteriores se suelen automatizar con la ayuda de algún programa estadístico o de análisis de datos. La figura 19 muestra cómo se pueden calcular probabilidades de una binomial con ayuda de Minitab. La figura 20, por su parte, muestra cómo obtenerlas usando Excel.

Figura 19. Cálculo de probabilidades en una binomial con Minitab y Excel

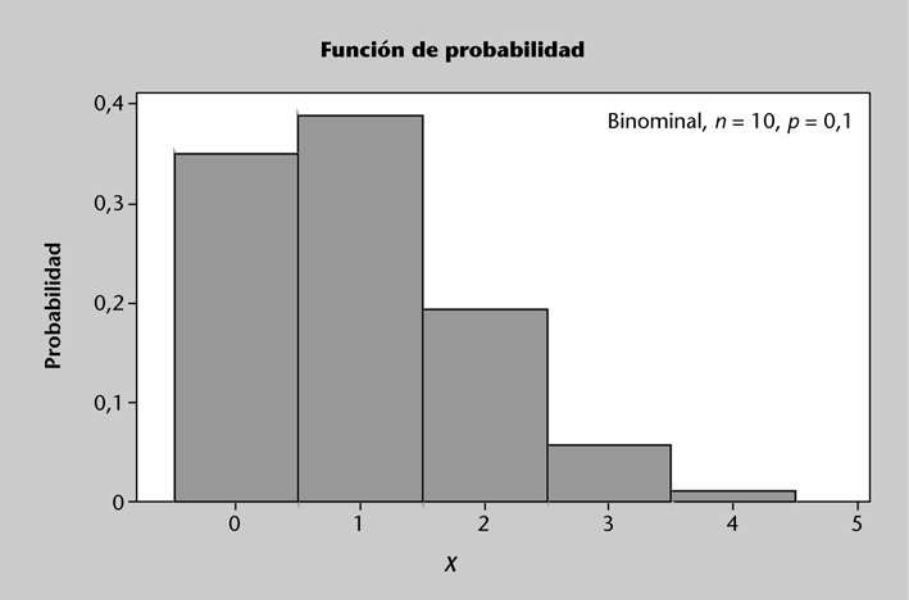


C1 =DISTR.BINOM(1,10,0.1,VERDADERO)

	A	B	C	D	E
1		P(X <= 1)	0,73609893		
2		P(X = 1)	0,38742049		
3		P(X = 0)	0,34867844		

La figura 20 se muestra la función de probabilidad asociada a la binomial del ejemplo anterior. Se observa que, aunque en teoría los posibles valores de la variable X irían desde 0 hasta 10 (número de pruebas), en la práctica los valores mayores de 4 tienen probabilidad de suceso prácticamente nula (por ejemplo, es muy poco frecuente que se obtengan valores superiores a 4). En efecto, $P(X > 4) = 1 - P(X <= 4) =$ {usando Minitab o Excel} $= 1 - 0{,}9984 = 0{,}0016$.

Figura 20. Función de probabilidad de una B (10, 0,1)

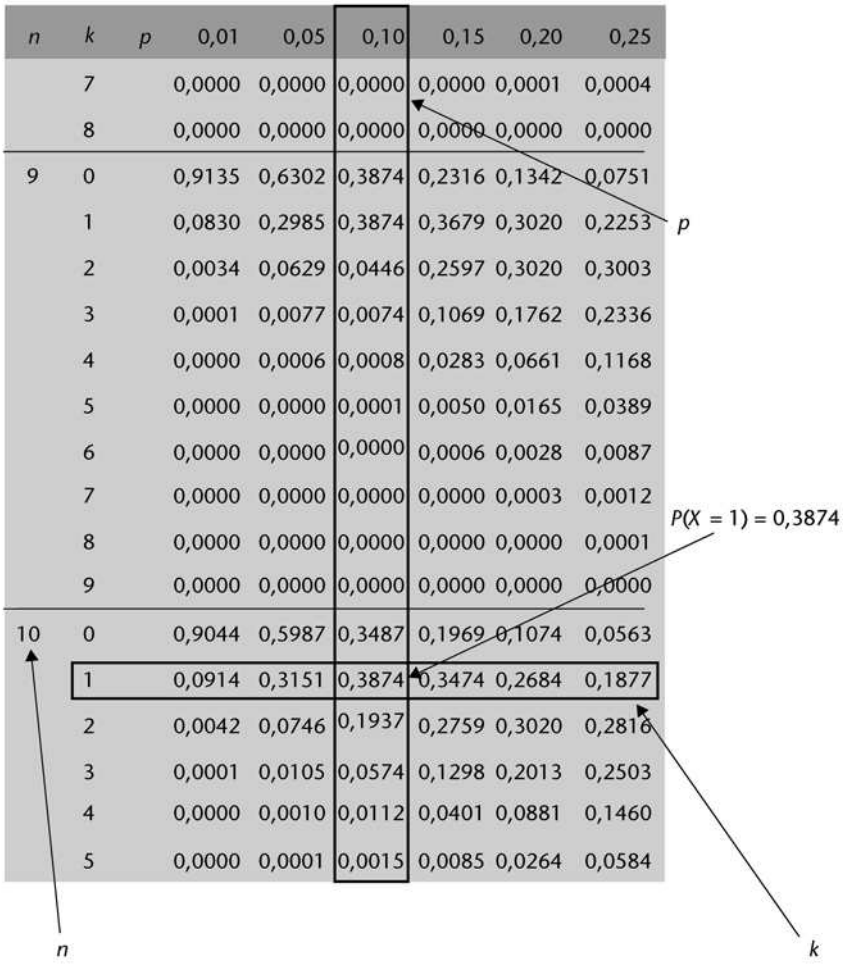


Las probabilidades anteriores se pueden obtener también mediante el uso de tablas estadísticas (sin necesidad de usar ningún software). Así, siguiendo el ejemplo anterior, la figura 21 muestra cómo calcular $P(X = 1)$ usando la tabla binomial. En este caso, X es una $B(10, 0{,}1)$ y se quiere hallar $P(X = k)$ siendo $k = 1$. Para ello, se busca la sección de la tabla correspondiente a $n = 10$, y la intersección entre la fila $k = 1$ y la columna $p = 0{,}1$.

Cálculo de probabilidades

Resulta fácil encontrar en Internet abundantes documentos que explican con todo detalle el uso de tablas para calcular probabilidades. En la medida de lo posible, sin embargo, conviene automatizar los cálculos mediante el uso de software.

Figura 21. Cálculo de probabilidades binomiales mediante tablas



n	k	p	0,01	0,05	0,10	0,15	0,20	0,25
	7		0,0000	0,0000	0,0000	0,0000	0,0001	0,0004
	8		0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
9	0		0,9135	0,6302	0,3874	0,2316	0,1342	0,0751
	1		0,0830	0,2985	0,3874	0,3679	0,3020	0,2253
	2		0,0034	0,0629	0,0446	0,2597	0,3020	0,3003
	3		0,0001	0,0077	0,0074	0,1069	0,1762	0,2336
	4		0,0000	0,0006	0,0008	0,0283	0,0661	0,1168
	5		0,0000	0,0000	0,0001	0,0050	0,0165	0,0389
	6		0,0000	0,0000	0,0000	0,0006	0,0028	0,0087
	7		0,0000	0,0000	0,0000	0,0000	0,0003	0,0012
	8		0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
	9		0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
10	0		0,9044	0,5987	0,3487	0,1969	0,1074	0,0563
	1		0,0914	0,3151	0,3874	0,3474	0,2684	0,1877
	2		0,0042	0,0746	0,1937	0,2759	0,3020	0,2816
	3		0,0001	0,0105	0,0574	0,1298	0,2013	0,2503
	4		0,0000	0,0010	0,0112	0,0401	0,0881	0,1460
	5		0,0000	0,0001	0,0015	0,0085	0,0264	0,0584

6. Distribuciones de probabilidad continuas

Al inicio de este módulo se definió el concepto de variable cuantitativa continua como aquella variable cuantitativa que podía tomar un número infinito (no contable) de valores distintos. Así, un ejemplo de variable continua sería X = "tiempo que se tarda en desarrollar un portal web", ya que esta variable puede tomar un valor real cualquiera entre 0 e infinito.

A diferencia de lo que ocurría con las variables discretas, cuando se trabaja con variables continuas no es posible definir una función de probabilidad que asigne probabilidades a los distintos valores de la variable: si X es una variable continua, X puede tomar un número infinito (no contable) de valores, por lo que la probabilidad teórica de que la variable X tome un valor concreto x es siempre 0, es decir: $P(X = x) = 0$ para cualquier valor x de X. Sí es posible, sin embargo, asignar probabilidades a intervalos de valores. Por ejemplo, si el 51% de los portales web tardan en desarrollarse entre 240 y 258 horas, entonces $P(240 < X < 258) = 0{,}51$. Para describir la distribución de probabilidad de una variable continua se sigue usando la función de distribución (aunque con algún matiz nuevo) y, además, se usa también la llamada "función de densidad" en lugar de la función de probabilidad típica de variables discretas:

Nota

En variables continuas, puesto que $P(X = x) = 0$ para cualquier valor x de X, se cumplirá que:

a) $P(X \leq x) = P(X < x)$

b) $P(X \geq x) = P(X > x)$

La **función de densidad** de una variable continua X es una función $f(x)$ tal que la probabilidad de que X tome un valor en un intervalo (a, b) coincide con el **área** "encerrada" por dicha función entre los extremos de dicho intervalo (figura 22), es decir: $P(a < X < b)$ = área bajo $f(x)$ entre a y b.

La **función de distribución** de X es aquella función $F(x)$ que asigna a cada posible valor x de X su probabilidad acumulada de ocurrencia (figura 23), es decir, $F(x) = P(X \leq x)$ = área bajo $f(x)$ desde $-\infty$ (menos infinito) hasta x.

Nota

La función de densidad $f(x)$ siempre es positiva y "encierra" un área total de 1.

Atención

Observar la equivalencia entre los conceptos de "probabilidad" y "área".

La figura 22 muestra la función de densidad de una variable con distribución simétrica y centrada en el valor 250 (puesto que la función es totalmente simétrica la media y la mediana coinciden en este punto). Se observa también el área encerrada bajo función de densidad entre los valores $a = 240$ y $b = 258$. Esta área corresponde con la probabilidad siguiente: $P(240 < X < 258)$. Por su parte, la figura 23 muestra la función de distribución asociada a la misma variable. Nuevamente se aprecia la simetría con respecto al valor central, así como el hecho de que la función de distribución va creciendo conforme va acumulando probabilidades, pasando del valor 0 en su extremo izquierdo al valor 1 en su extremo derecho. A partir de esta gráfica se pueden estimar visualmente probabilidades acumuladas, por ejemplo: $P(X <= 260)$ será un valor muy cercano a 0,8.

Figura 22. Función de densidad de una variable continua y área encerrada

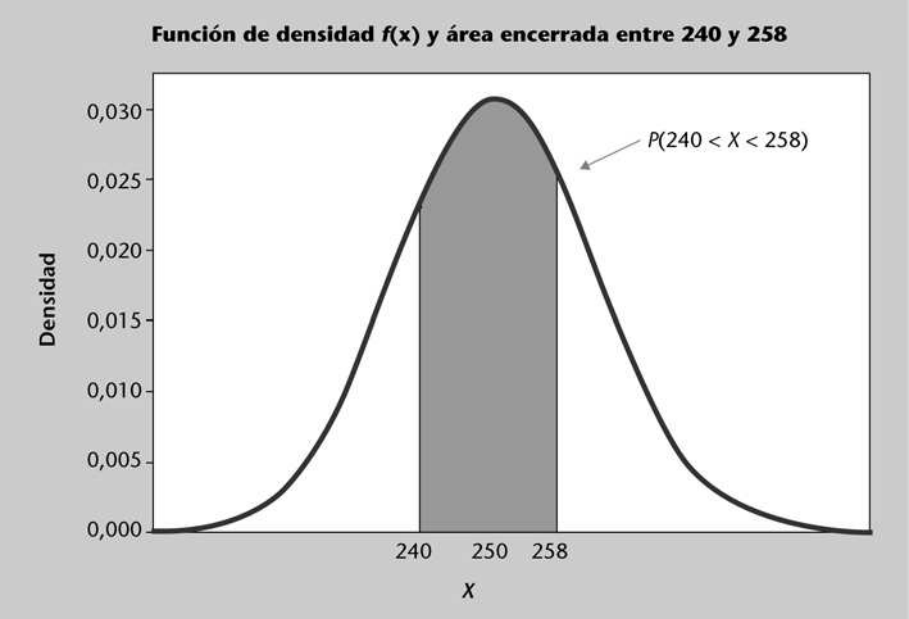
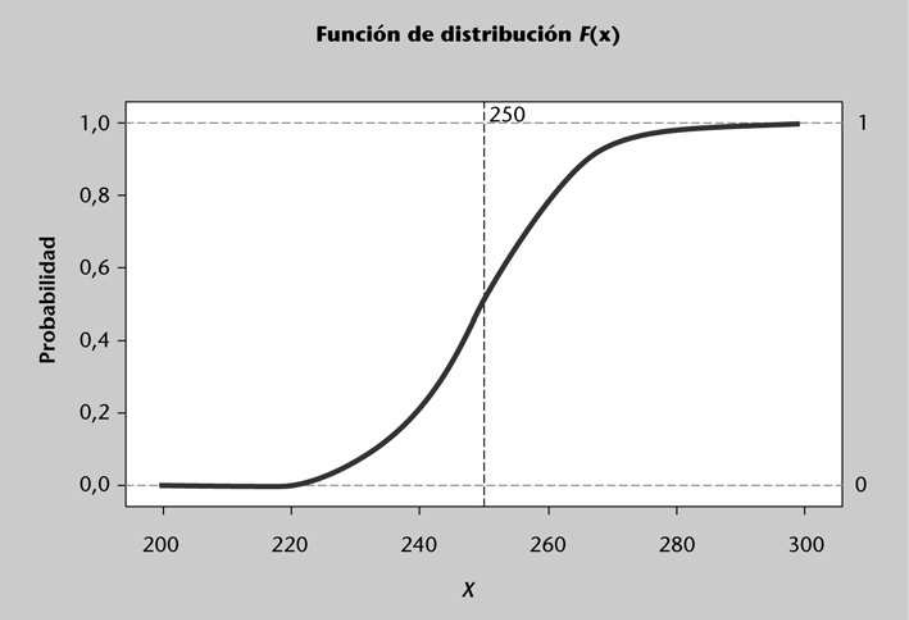


Figura 23. Función de distribución de una variable continua



Función de distribución

La función de distribución es una función acumulativa de probabilidades y, por tanto, es siempre creciente, pasando de 0 (extremo izquierdo) a 1 (extremo derecho).

Parámetros descriptivos de una distribución continua

En el caso de distribuciones continuas, la forma de calcular los parámetros es similar a la empleada para distribuciones discretas, si bien ahora los sumatorios se sustituyen por áreas (integrales definidas en términos matemáticos) entre dos extremos:

Atención

Aunque en la práctica se hará uso de programas estadísticos para hacer los cálculos, es importante conocer qué conceptos se usan para definir cada tipo de parámetro.

- **Media o valor esperado de una variable continua:** la media o valor esperado de una variable continua X se representa por μ o $E[X]$ y se calcula de la siguiente forma:

$$\mu = E[X] = \text{área total bajo } "x \cdot f(x)" = \int_{-\infty}^{+\infty} x \cdot f(x)dx$$

donde $f(x)$ denota a la función de densidad de X.

- **Varianza y desviación estándar de una variable continua:** la varianza de una variable continua X se representa por σ^2 y se calcula de la siguiente forma:

$$\sigma^2 = \text{área total bajo } "(x-\mu)^2 \cdot f(x)" = \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot f(x)\,dx$$

donde $f(x)$ denota a la función de densidad de X. Como siempre, la desviación estándar de una variable es la raíz cuadrada positiva de su varianza, es decir:

$$\sigma = \sqrt{\sigma^2}$$

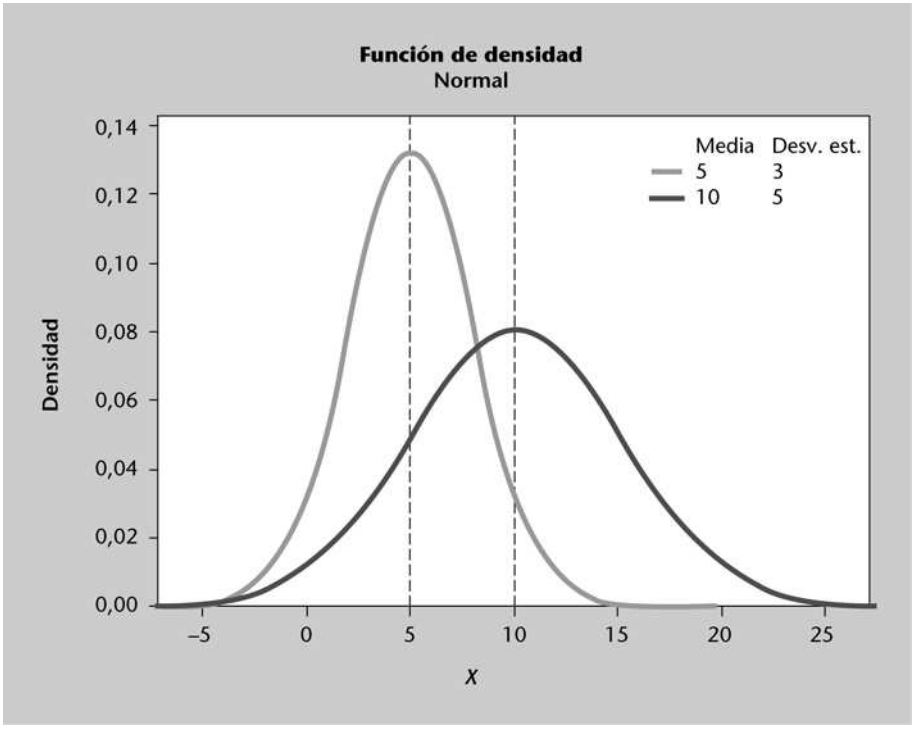
La distribución normal o gaussiana

La distribución normal o gaussiana es la distribución teórica más importante. Muchas variables continuas siguen una distribución normal o aproximadamente normal. Otras variables continuas y discretas también pueden, en determinadas circunstancias, ser aproximadas mediante una distribución normal. La normal, además, es una distribución clave en la estadística inferencial ya que algunas de sus propiedades se utilizan para obtener información sobre toda la población a partir de información sobre una muestra.

La forma concreta de una distribución normal viene caracterizada por dos parámetros: la media, μ, que define dónde se sitúa el centro de la función de densidad, y la desviación estándar, σ, que define la amplitud de la función de densidad. Cuando una variable continua X sigue una distribución normal, se suele representar por $X \sim N(\mu, \sigma)$.

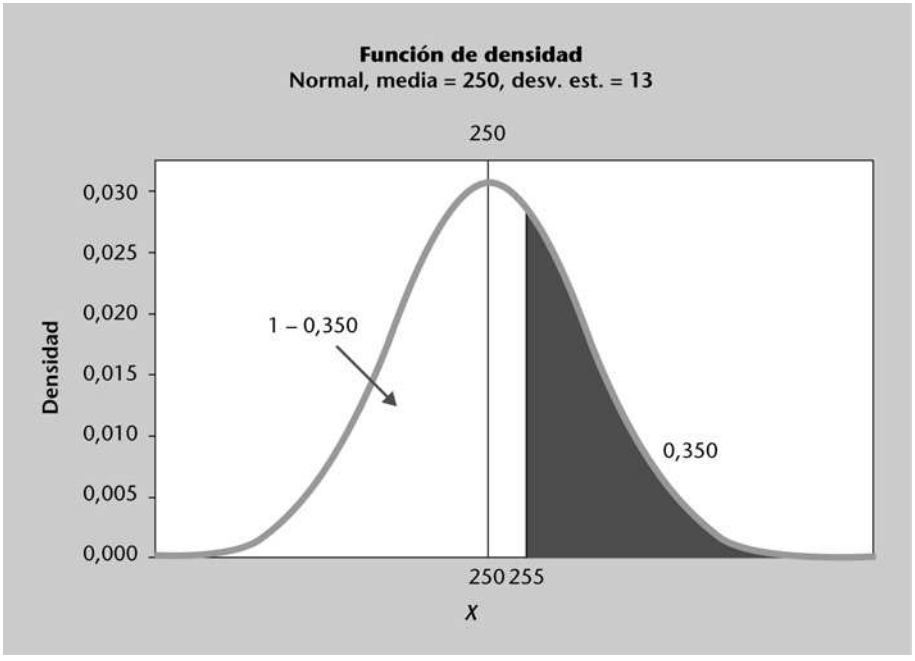
Las figuras 22 y 23 muestran, respectivamente, la función de densidad y la función de distribución de una normal con media $\mu = 250$ y desviación estándar $\sigma = 13$. La figura 24 muestra las funciones de densidad para dos distribuciones de tipo normal con parámetros $\{\mu = 5, \sigma = 3\}$ y $\{\mu = 10, \sigma = 5\}$ respectivamente. Se observa que la función de densidad de la normal tiene forma de "campana de Gauss", elevada en el centro (el valor medio o esperado) y con dos colas simétricas en los extremos. Es de destacar, además, cómo cada una de las curvas está centrada en su media, así como el hecho de que la curva es más ancha cuanto mayor es la desviación estándar.

Figura 24. Funciones de densidad asociadas a sendas normales



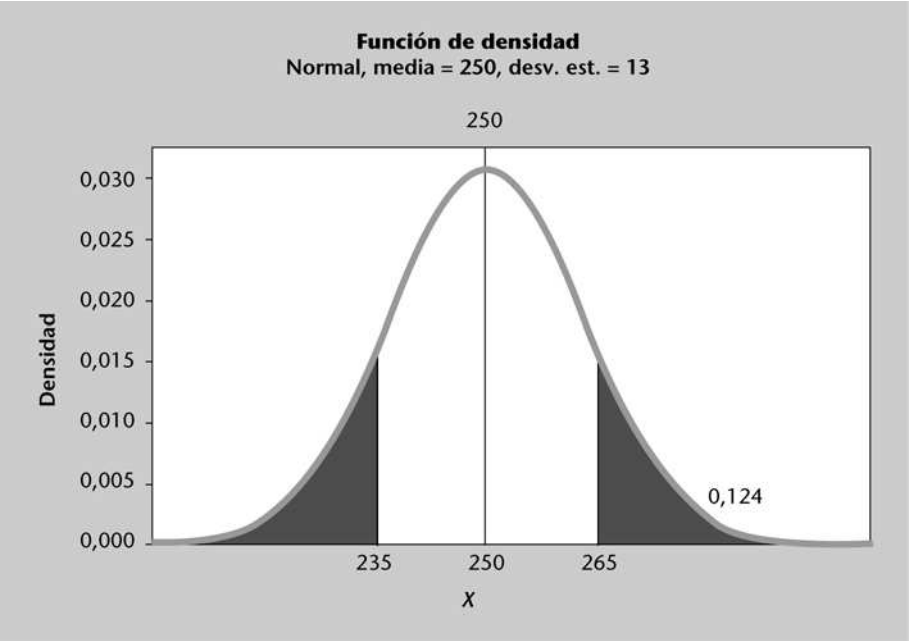
Como en cualquier otra función de densidad, el área total encerrada bajo la curva es de 1. En la práctica eso significa que para cualquier valor x de X, $P(X > x) = 1 - P(X < x)$, es decir, el área a la derecha de un valor es el área total (que vale 1) menos el área a su izquierda y viceversa (figura 25). Además, puesto que la normal es una distribución simétrica con respecto a su media, el área "encerrada" por una cola es igual al área "encerrada" por la cola opuesta (figura 26).

Figura 25. El área total de una función de densidad es 1



Cualquier distribución normal cumple además la llamada **regla 68-95-99,7** según la cual el intervalo $(\mu - \sigma, \mu + \sigma)$ contiene aproximadamente el 68% de las observaciones, el intervalo $(\mu - 2\sigma, \mu + 2\sigma)$ contiene aproximadamente el 95% de las observaciones y el intervalo $(\mu - 3\sigma, \mu + 3\sigma)$ contiene aproximadamente el 99,7% de las observaciones. Así, por ejemplo, si $X \sim N(250,13)$ se puede afirmar que un 68% de las observaciones de X estarán en el intervalo (237, 263), un 95% de las observaciones estarán en el intervalo (224, 276) y un 99,7% de las observaciones estarán en el intervalo (211, 289). Observad, por tanto, que será altamente improbable encontrar valores de X fuera de este último intervalo.

Figura 26. Dos colas simétricas "encierran" la misma área



De entre las infinitas distribuciones normales que se pueden considerar variando los parámetros μ y σ conviene citar la llamada **normal estándar**, que tiene por parámetros $\mu = 0$ y $\sigma = 1$. En otras palabras, una variable continua Z se distribuirá según una normal estándar, $Z \sim N(0,1)$, si su función de densidad es la de una normal centrada en el origen y con desviación estándar unitaria. Esta distribución normal estándar se suele usar bastante en estadística inferencial y también cuando se desean calcular probabilidades de una normal cualquiera mediante el uso de tablas de probabilidades ya calculadas.

En efecto, dada una variable normal cualquiera, $X \sim N(\mu, \sigma)$, es posible aplicarle un **proceso de estandarización** para obtener una normal estándar Z. Esto se consigue restando a la variable X su media μ (con lo que la función de densidad se desplaza a lo largo del eje x hasta que queda centrada en el origen) y dividiendo el resultado por su desviación estándar σ (con lo que la nueva variable tendrá una desviación estándar unitaria), es decir: $Z = \frac{X - \mu}{\sigma} \sim N(0,1)$. Este proceso de estandarización permite, entre otras cosas, calcular probabilidades para una normal cualquiera a partir de las tablas de probabilidades precalculadas que existen para la distribución

normal estándar, lo que evita el tener que resolver integrales cada vez que se desea obtener una nueva probabilidad. Supongamos, por ejemplo, que X sigue una $N(1.500, 100)$ y se desea obtener $P(X < 1.400)$ mediante el uso de tablas. El primer paso consiste en estandarizar los valores:

$$P(X<1.400)=P\left(\frac{X-\overline{x}}{\sigma}<\frac{1.400-\overline{x}}{\sigma}\right)=P\left(Z<\frac{1.400-1.500}{100}\right)=P\left(Z<-1\right)$$

En otras palabras, se desea calcular el área a la izquierda del valor –1 en una normal tipificada o estándar. Normalmente, la tabla de la normal estándar, Z, ofrece áreas (probabilidades) a la izquierda de valores positivos, por lo que resultará necesario hacer una pequeña transformación teniendo en cuenta que: (a) por simetría de la normal estándar, el área (probabilidad) a la izquierda de un valor negativo k es igual al área (probabilidad) a la derecha del correspondiente valor positivo, $|k|$ (p. ej., $P(Z < -1) = P(Z > 1)$), y (b) el área (probabilidad) total encerrada bajo la curva es 1 (p. ej., el área a la izquierda de un valor más el área a su derecha suma 1, por ejemplo: $P(Z < 1) + P(Z > 1) = 1$). Teniendo en cuenta lo anterior, se deduce que $P(Z < -1) = P(Z > 1) = 1 - P(Z < 1) =$ {ver tabla figura 27} $= 1 - 0,8413 = 0,1587$.

Figura 27. Cálculo de probabilidades en una normal mediante tablas

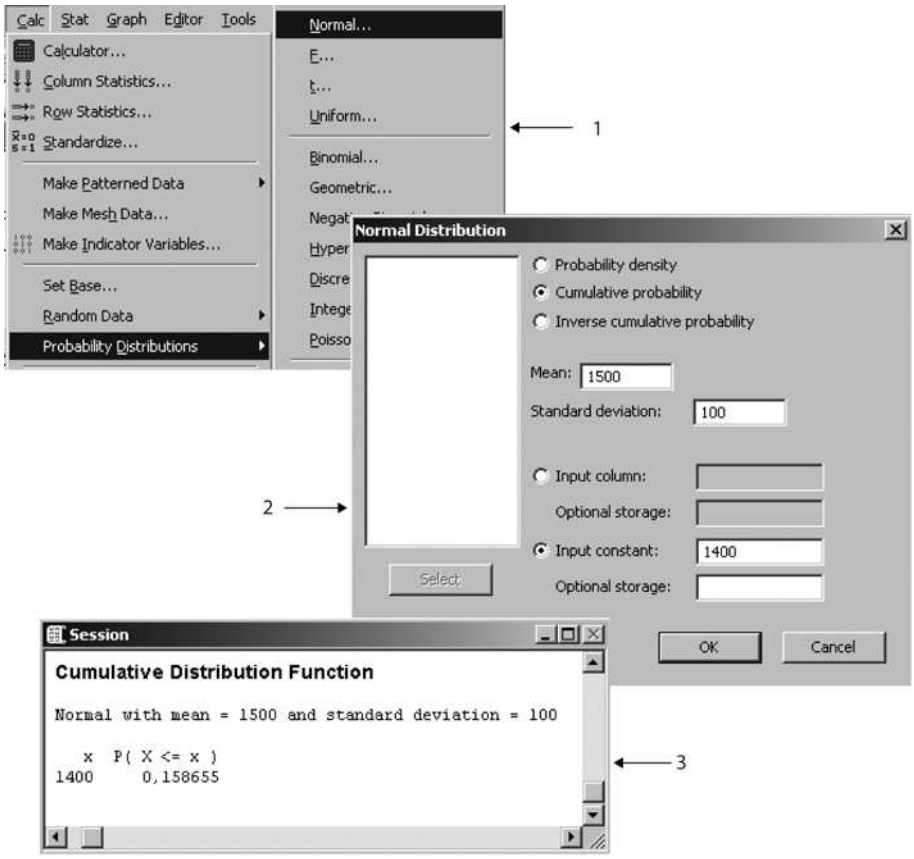
	,00	,01	,02	,03	,04	,05
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265

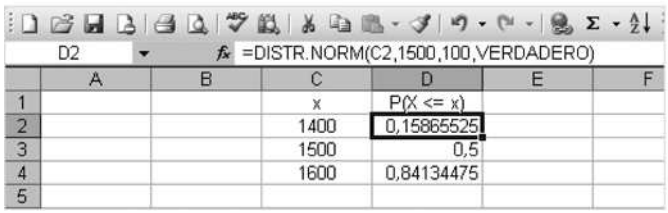
Nota

Notar que para hallar $P(Z < 1,00)$ usando la tabla se ha de buscar el valor intersección entre la fila 1,0 y la columna 0,00 (dado que 1,00 = 1,0 + 0,00). Si se pidiese $P(Z < 1,24)$, entonces habría que buscar la intersección entre la fila 1,2 y la columna 0,04 (dado que 1,24 = 1,2 + 0,04), con lo que se obtendría el valor 0,8925.

Por otra parte, también es posible automatizar el cálculo de probabilidades de una normal cualquiera mediante el uso de programas estadísticos, con lo que se elimina así la necesidad de resolver manualmente las integrales indefinidas o de tener que usar tablas de probabilidades precalculadas. La figura 28 muestra cómo obtener probabilidades de una normal con Minitab. En concreto, para una normal con media $\mu = 1.500$ y desviación estándar $\sigma = 100$, se obtiene que $P(X < 1.400) = 0,158655$. Asimismo, la figura 28 muestra cómo se han obtenido con Minitab y Excel algunas probabilidades para la misma variable. Es preciso observar que $P(X < 1.500) = 0,5$, lo cual es lógico puesto que 1.500 es la media y, a la vez, la mediana de la distribución normal.

Figura 28. Cálculo de probabilidades en una normal con Minitab y Excel





Pasos a seguir

Se sigue la ruta ***Calc*** > ***Probability Distributions*** > ***Normal (1)*** y se completan los parámetros en la ventana correspondiente ***(2)***. El resultado se muestra en ***(3)***. Observar que, si en lugar de escoger la opción ***Cumulative probability*** en ***(2)*** se hubiera escogido la opción ***Probability density,*** el programa hubiera calculado el valor de la función de densidad en $x = 1.400$ en lugar de $P(X < 1.400)$. Finalmente, para una probabilidad p dada, la opción ***Inverse cumulative probability*** devuelve aquel valor c de la variable X tal que $P(X < c) = p$.

Ejemplos de aplicación de una normal

- Según un estudio realizado por el Ministerio de Educación, el número de horas anuales que dedican los niños españoles a ver la televisión es una variable aleatoria que sigue una distribución normal de media 1.500 horas y desviación estándar de 100 horas. ¿Qué porcentaje de niños dedican entre 1.400 y 1.600 horas anuales?

  En este caso, $X \sim N$ (1.500,100) y se pide $P(1.400 < X < 1.600)$. Por la regla 68-95-99,7, se tiene que la probabilidad anterior será, aproximadamente, del 68% (ya que $\mu - \sigma = 1.400$ y $\mu + \sigma = 1.600$). Para calcular de forma más exacta dicha probabilidad, conviene notar que $P(1.400 < X < 1.600) = P(X < 1.600) - P(X < 1.400)$, es decir: el área entre 1.400 y 1.600 coincide con el área a la izquierda de 1.600 menos el área a la izquierda de 1.400. Las probabilidades anteriores se pueden calcular usando cualquier programa estadístico (p. ej.: Minitab o Excel), y resultan: $P(X < 1.600) = 0,8413$ y $P(X < 1.400) = 0,1587$, por lo que la probabilidad buscada es de 0,6827, es decir, un 68,27% de los niños dedican entre 1.400 y 1.600 horas anuales a ver la televisión.

- En base a los datos del Instituto Nacional de Estadística (INE), el sueldo medio anual de un trabajador es de 26.362 euros. Suponiendo que dichos sueldos sigan una distribución normal con una desviación estándar de 6.500 euros, ¿cuál será el porcentaje de trabajadores que superen los 40.000 euros?

  En este caso, $X \sim N$ (26.362,6.500) y se pide $P(X > 40.000)$. Observar que, puesto que el área total bajo la curva normal es 1, $P(X > 40.000) = 1 - P(X < 40.000)$ = {Minitab o Excel} = 1 – 0,9821 = 0,0179, es decir, sólo un 1,8% de los trabajadores superarían la cifra de los 40.000 euros anuales.

- El tiempo que se emplea en rellenar un cuestionario en línea sigue una distribución aproximadamente normal con una media de 3,7 minutos y una desviación estándar de 1,4 minutos. ¿Cuál es la probabilidad de que se tarde menos de 2 minutos en responder a dicho cuestionario? ¿Y de que se tarde más de 6 minutos? Hallad el valor c tal que $P(X < c) = 0,75$ (percentil 75 de la variable).

  En este caso, $X \sim N$ (3,7, 1,4). En primer lugar, $P(X < 2)$ = {Minitab o Excel} = 0,1131, es decir: un 11,31% de los individuos que respondan el cuestionario emplearan menos de 2 minutos en hacerlo. Por otra parte, $P(X > 6) = 1 - P(X < 6)$ = {Minitab o Excel} = 0,0505, es decir, un 5% de los individuos tardarán más de 6 minutos en responder el cuestionario. Finalmente, para hallar el valor c tal que $P(X < c) = 0,75$ se debe usar la opción *Inverse cumulative probability* de Minitab (o su equivalente en Excel), con lo que se obtiene un valor aproximado de 4,64 minutos, es decir el 75% de los individuos tardan menos de 4,64 minutos en completar el cuestionario (o, dicho de otro modo, el 25% tardan más de 4,64 minutos en hacerlo).

Las distribuciones *t*-Student y *F*-Snedecor

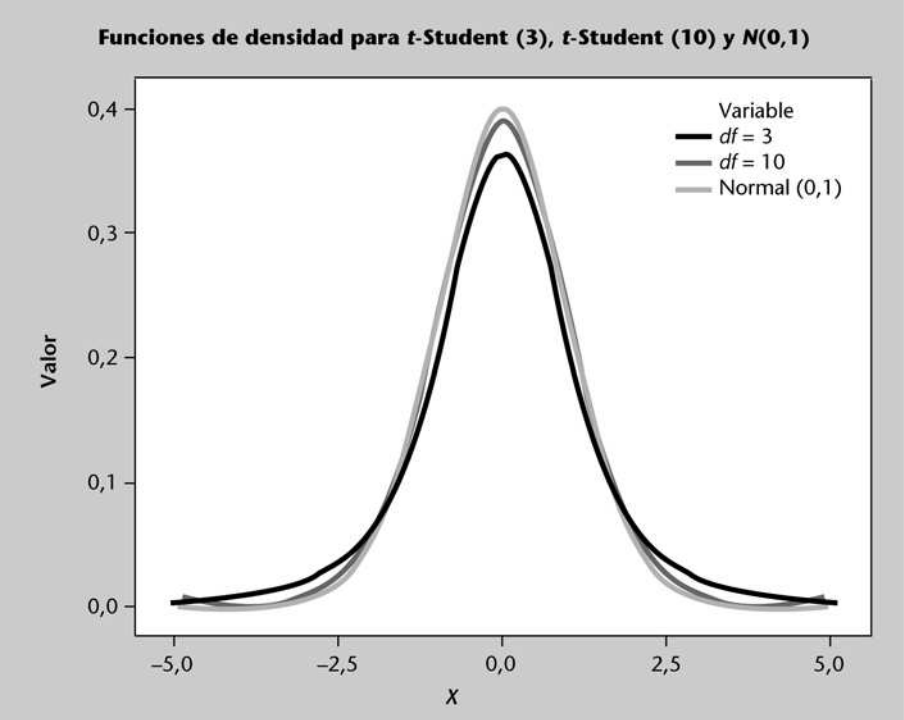
Además de la normal, hay muchas otras distribuciones de probabilidad continuas que se suelen usar en estadística inferencial. Una de ellas es la llamada distribución *t*-Student, y otra es la llamada *F*-Snedecor. Ambas se presentan a continuación:

La distribución ***t*-Student** es una distribución simétrica y centrada en el origen (es decir, su media y su mediana son 0). Esta distribución se caracteriza por un parámetro llamado **grados de libertad** o ***df*** (*degrees of freedom*), siendo $df > 2$. En la práctica, $df = n - 1$, donde n es el tamaño de la muestra que se esté analizando. La figura 29 muestra diversas funciones de densidad de las *t*-Student, cada una de ellas asociadas a un valor concreto del parámetro *df*. Se observa cómo la *t*-Student se asemeja cada vez más a una normal estándar conforme se va incrementando el parámetro grados de libertad.

Grados de libertad

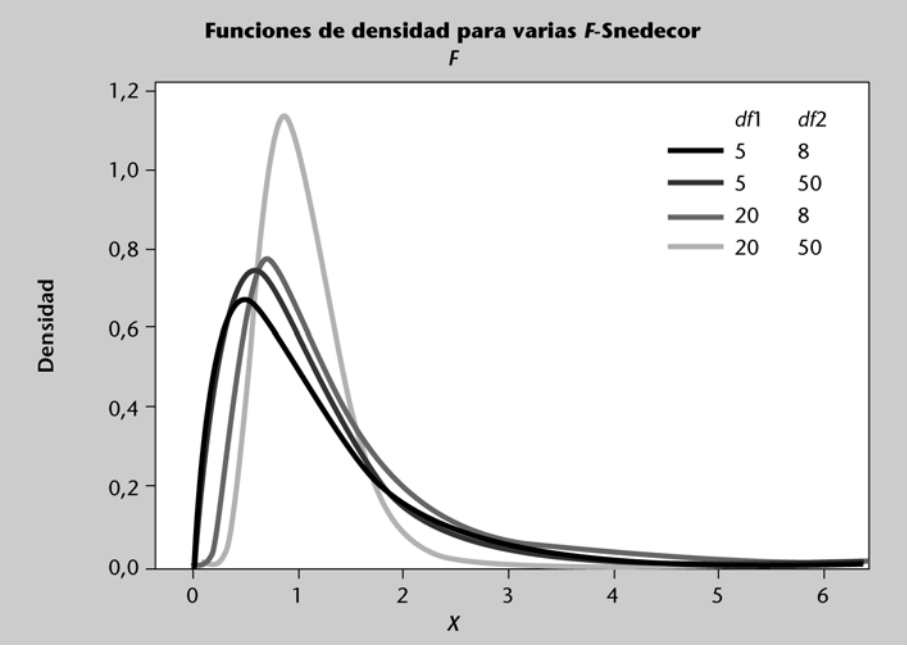
En estadística, el concepto de **grados de libertad** asociados a un conjunto de datos se puede interpretar como el número mínimo de valores que se necesitaría conocer para determinar dichos datos. Así, por ejemplo, en el caso de un muestra aleatoria de tamaño N, habría N grados de libertad (no se puede determinar el valor de ninguno de los datos incluso aunque se conociese el valor de los $N - 1$ restantes). Sin embargo, un conjunto de N datos de los cuales se conozcan $N - 1$, la media muestral tendría $N - 1$ grados de libertad (fijados los valores de los $N - 1$ datos y de la media, quedaría ya fijado el valor desconocido restante). Así, si tenemos un conjunto de 3 observaciones de la variable X, $x_1 = 2$, $x_2 = -2$ y $x_3 = a$ (desconocido), y sabemos que la media de los tres valores es 0, necesariamente $a = 0$.

Figura 29. Funciones de densidad de *t*-Student según *df*



Por su parte, la distribución ***F*-Snedecor** es otra distribución continua. La *F*-Snedecor siempre toma valores no negativos (es decir, una variable que siga dicha distribución sólo puede tomar valores iguales o mayores a 0, nunca valores negativos). Además, esta distribución no es simétrica, sino que está sesgada a la derecha (figura 30). Así como la normal venía caracterizada por dos parámetros, μ (media) y σ (desviación estándar), la *F*-Snedecor también se caracteriza por dos parámetros: los **grados de libertad del numerador**, *df*1 y los **grados de libertad del denominador**, *df*2. Al igual que ocurría con la *t*-Student, para cada valor de estos parámetros se obtiene una función de densidad distinta y, por tanto, una distribución *F*-Snedecor distinta.

Figura 30. Funciones de densidad de *t*-Student según *df*1 y *df*2

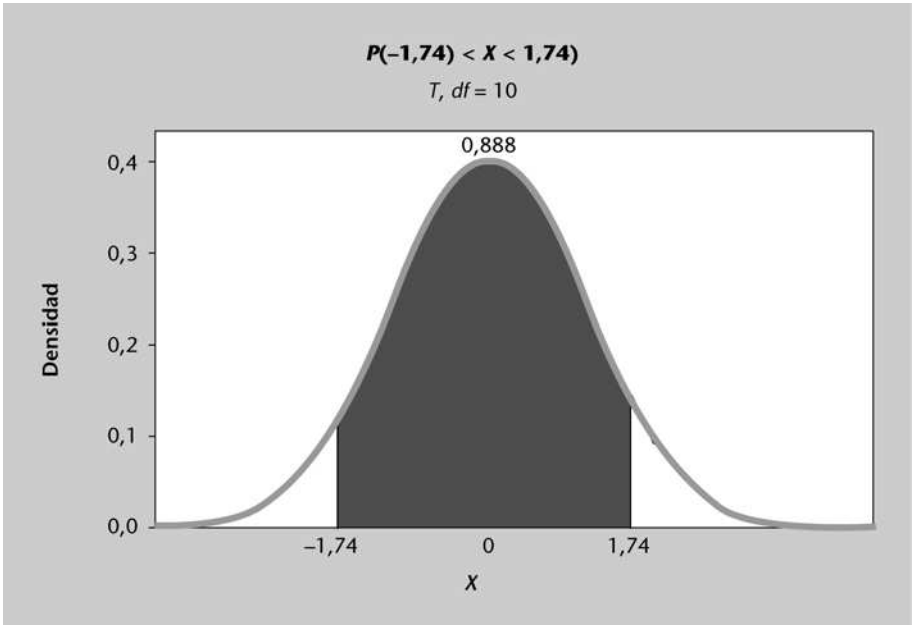


Para calcular probabilidades asociadas a una *t*-Student o a una *F*-Snedecor, pueden usarse programas estadísticos o de análisis de datos (Minitab, Excel, etc.) de forma análoga a como se hacía en el caso de la normal. Así, por ejemplo, si X es una variable aleatoria que sigue una distribución *t*-Student con diez grados de libertad, $P(-1{,}74 < X < 1{,}74) = P(X < 1{,}74) - P(X < -1{,}74) =$ {Minitab o Excel} $= 0{,}9438 - 0{,}0562 = 0{,}8876$ (figura 31).

Nota

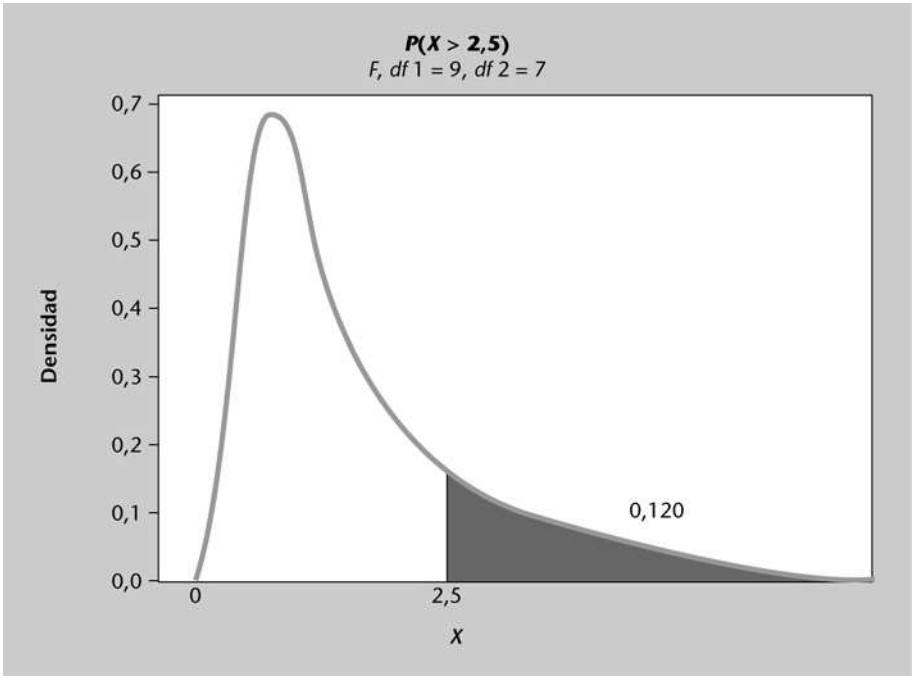
Notar que $P(-1{,}74 < X < 1{,}74)$ viene representada por el área marcada en la figura 31 (esto es, el área comprendida entre los valores –1,74 y 1,74). Para calcular dicha área, se calcula $P(X < 1{,}74)$ (p. ej., el área a la izquierda del 1,74) y al valor obtenido se le resta $P(X < -1{,}74)$ (p. ej., el área a la izquierda del -1,74). Para calcular $P(X < 1{,}74)$ con Minitab se usa el menú ***Calc*** > ***Probability Distributions*** > ***t…***, especificando los grados de libertad (10 en este ejemplo) y el valor de la constante (1,74 en este caso). Análogamente se obtendría el valor de $P(X < -1{,}74)$.

Figura 31. Probabilidades en una *t*-Student



Finalmente, si X es una variable aleatoria que sigue una distribución *F*-Snedecor con nueve grados de libertad en el numerador y siete grados de libertad en el denominador, entonces $P(X > 2{,}5) = 1 - P(X < 2{,}5) =$ {Minitab o Excel} $= 1 - 0{,}8797 = 0{,}1203$ (figura 32).

Figura 32. Probabilidades en una *F*-Snedecor



Nota

De forma análoga a como ocurría en el caso de las distribuciones binomial y normal, también existen tablas que permiten calcular, sin necesidad de utilizar software como Minitab o Excel, las probabilidades asociadas a una distribución *t*-Student o *F*-Snedecor (ver, p. ej., http://www.statsoft.com/textbook/distribution-tables).

Resumen

En este módulo se han presentado las técnicas básicas de la estadística descriptiva univariante: representación gráfica de datos discretos y continuos, organización de los datos mediante tablas de frecuencias y uso de estadísticos descriptivos para resumir datos. Conviene recordar que el tipo de gráfico, tabla o estadístico a usar dependerá siempre del tipo de variable considerada (categórica, cuantitativa discreta o cuantitativa continua), así como del tipo de información que se desee obtener.

Además, se ha explicado también el concepto de probabilidad de un suceso, que desempeña una función relevante en el análisis y predicción del comportamiento de las variables aleatorias asociadas a fenómenos cotidianos.

Finalmente, se han presentado algunos de los principales modelos matemáticos que se usan para describir, de forma teórica, el comportamiento de variables aleatorias: la distribución binomial, la normal, la *t*-Student y la *F*-Snedecor son algunos ejemplos de dichos modelos. El cálculo de probabilidades asociadas a variables que se comportan según alguno de estos modelos permite entender mejor su comportamiento y realizar estimaciones sobre la población de individuos de la que provienen los datos.

Ejercicios de autoevaluación

1) La tabla siguiente resume las respuestas ofrecidas por doscientos usuarios de un portal web a la pregunta "el nivel de usabilidad del portal es adecuado":

Respuesta	Frecuencia
Totalmente de acuerdo	50
De acuerdo	75
Ligeramente de acuerdo	25
Ligeramente en desacuerdo	15
En desacuerdo	15
Totalmente en desacuerdo	20

Se pide que hagáis lo siguiente:
a) Construir un diagrama de barras que permita visualizar las respuestas obtenidas.
b) Calcular la frecuencia relativa de aparición de cada respuesta y construir un diagrama circular para ilustrar dichos valores.

2) La tabla siguiente contiene cuarenta observaciones para el tiempo transcurrido (en horas) entre el envío de un mensaje a un foro en línea y su correspondiente respuesta.

4,0	3,5	3,1	6,0	5,6	3,1	2,9	3,8
4,3	3,8	4,5	3,5	4,5	6,1	2,8	5,0
5,4	3,8	6,8	4,9	3,6	3,6	3,8	3,7
4,1	2,0	3,7	5,7	7,8	4,6	4,8	2,8
5,0	5,2	4,0	5,4	4,6	3,8	4,0	2,9

A partir de estos datos, debéis hacer lo siguiente:
a) Construir un diagrama de tallos y hojas. Usad 1,0 como unidad de incremento.
b) Construir un histograma.
c) ¿Se observa en los datos algún patrón claro? ¿Cuál es la moda de la distribución de los datos?

3) La tabla siguiente muestra veinte observaciones de la variable aleatoria "número de correos electrónicos recibidos en un día".

3,9	3,4	5,1	2,7	4,4
7,0	5,6	2,6	4,8	5,6
7,0	4,8	5,0	6,8	4,8
3,7	5,8	3,6	4,0	5,6

Se pide que hagáis lo siguiente:
a) Hallar los estadísticos descriptivos de esta muestra. ¿Cuánto vale el rango intercuartílico? ¿Entre qué dos valores están comprendidos el 50% de los datos centrales de la muestra?
b) Construir un diagrama de cajas y bigotes (*boxplot*). ¿Hay algún valor anómalo (*outlier*) entre las observaciones?

4) Cuando se efectúa un control antidopaje a un atleta que no ha tomado sustancia alguna, la probabilidad de que el test dé un falso positivo es de 0,006. Si durante una competición se efectúa el test a un total de 1.000 atletas que están libres de sustancias, ¿cuál será el número esperado (promedio) de falsos positivos?, ¿cuál es la probabilidad de que el número de falsos positivos sea superior a quince?, ¿qué cabría pensar si aparecen más de quince positivos?

5) De acuerdo con el Instituto Nacional de Estadística, el 9,96% de los adultos residentes en España son extranjeros. Con el fin de realizar una encuesta, se pretende contactar con una muestra aleatoria de mil doscientos adultos residentes en España. ¿Cuál será el número espe-

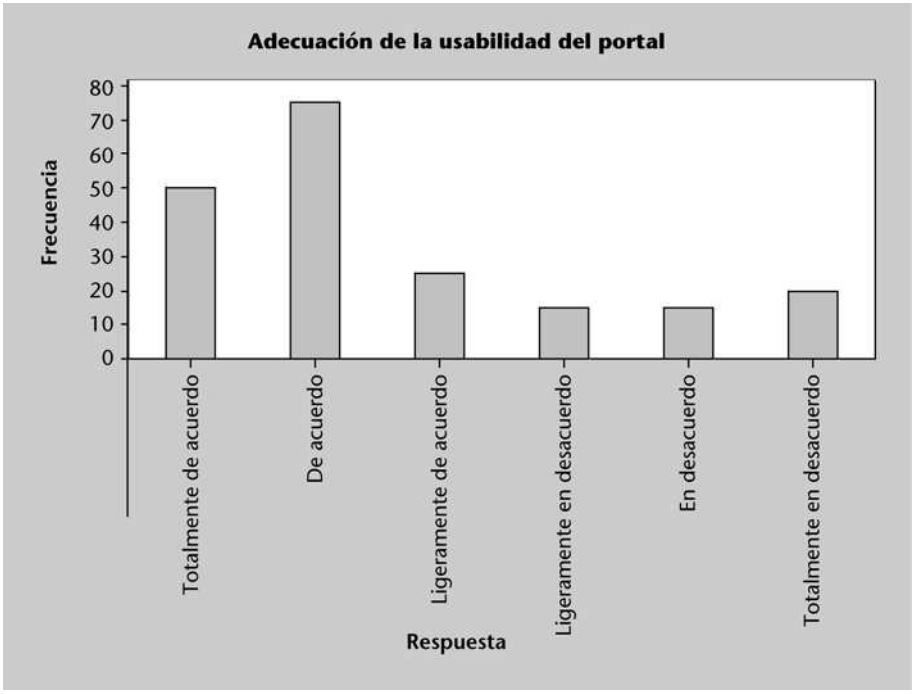
rado (promedio) de extranjeros que contendrá dicha muestra?, ¿cuál es la probabilidad de que la muestra contenga menos de cien extranjeros?

6) El tiempo de duración de un embarazo es una variable aleatoria que se distribuye de forma aproximadamente normal con una media de doscientos sesenta y seis días y una desviación estándar de dieciséis días. ¿Qué porcentaje de embarazos duran menos de doscientos cuarenta días (unos ocho meses)?, ¿qué porcentaje de embarazos duran entre doscientos cuarenta y doscientos setenta días (entre unos ocho y nueve meses)?, ¿a partir de cuántos días se sitúan el 20% de los embarazos más largos?

Solucionario

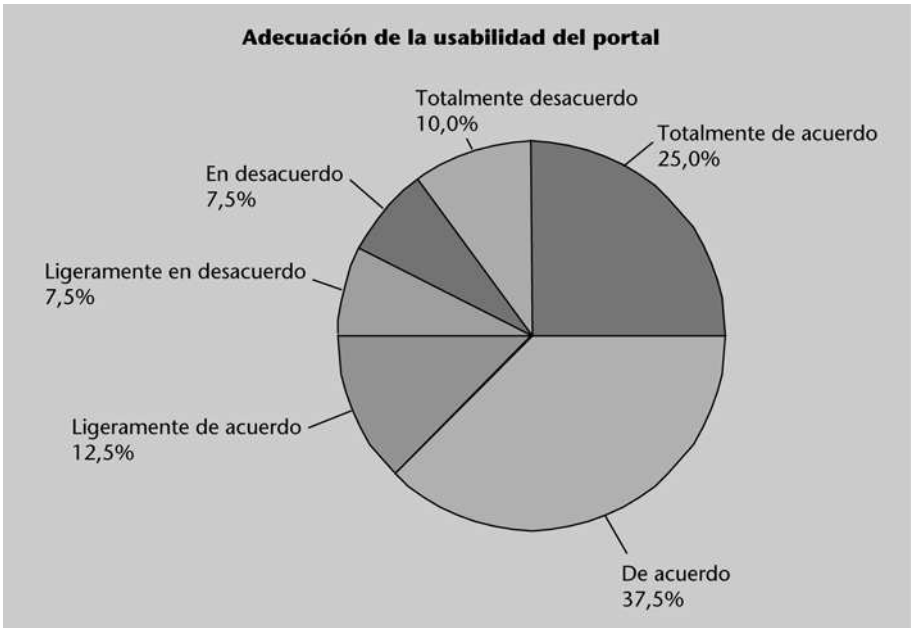
1)

a)



b)

Respuesta	Frecuencia	Frec. relativa
Totalmente de acuerdo	50	25,0%
De acuerdo	75	37,5%
Ligeramente de acuerdo	25	12,5%
Ligeramente en desacuerdo	15	7,5%
En desacuerdo	15	7,5%
Totalmente en desacuerdo	20	10,0%
Totales	**200**	**100%**



2)

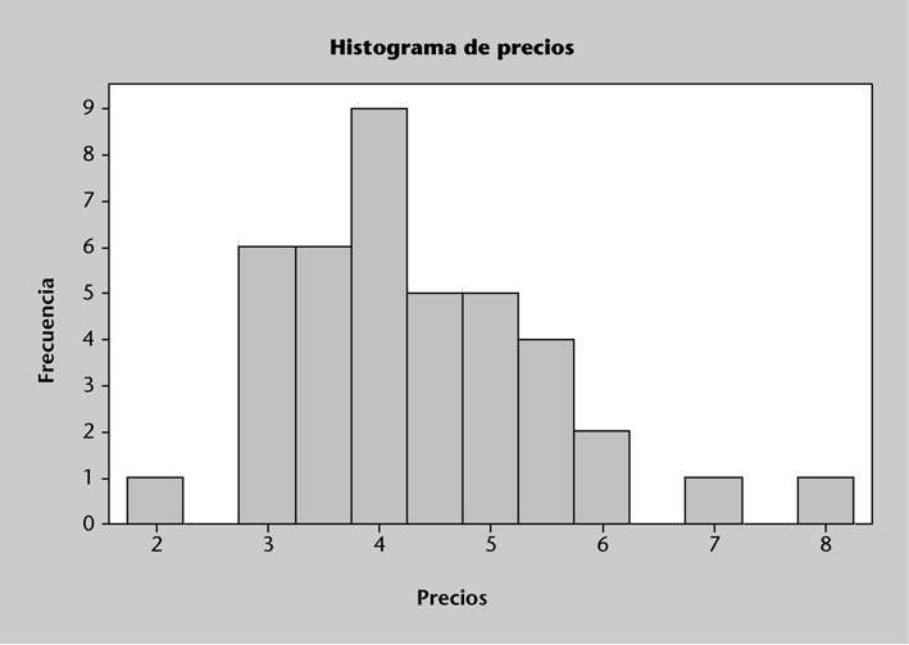
a)

Stem-and-Leaf Display: precios

```
Stem-and-leaf of precios  N  = 40
Leaf Unit = 0.10

 5    2  08899
 18   3  1155667788888
(11)  4  00013556689
 11   5  0024467
 4    6  018
 1    7  8
```

b)



c) Aunque no parece haber ningún patrón claro en los datos, sí se aprecia –tanto en el histograma como en el gráfico de tallos y hojas– una cierta forma de campana, con la parte central más elevada y unos extremos o colas más bajas. La moda de este conjunto de datos es 3,8 ya que, como se aprecia en el diagrama de tallos y hojas, es el valor que más aparece.

3)

a)

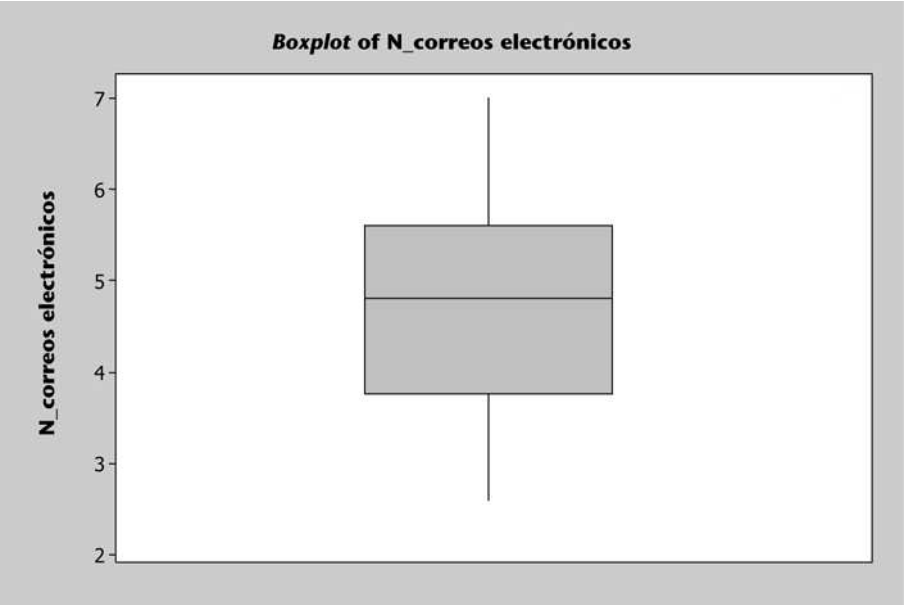
Descriptive Statistics: N_e-mails

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
N_e-mails	20	0	4.810	0.291	1.302	2.600	3.750	4.800	5.600

Variable	Maximum
N_e-mails	7.000

El rango intercuartílico es $Q3 - Q1 = 5{,}60 - 3{,}75 = 1{,}85$. Entre $Q1 = 3{,}75$ y $Q3 = 5{,}60$ están comprendidos el 50% de los datos centrales.

b)



No se observa, en este caso, ningún valor anómalo (*outlier*), ya que el gráfico no muestra ningún símbolo "*".

4) En este caso, puesto que el resultado de cada test puede ser "positivo" (con probabilidad 0,006) o "no positivo" (con probabilidad 1 – 0,006 = 0,994), la variable aleatoria X = "número de falsos positivos en 1.000 pruebas a atletas limpios" sigue una distribución binomial de parámetros n = 1.000 y p = 0,006. En el caso de la binomial, la media o valor esperado es $\mu = n \cdot p = 6$, es decir, cabe esperar que al aplicar el test a 1.000 atletas "limpios" haya seis falsos positivos.

Por otra parte, $P(X > 15) = 1 - P(X <= 15)$ = {Minitab o Excel} = 1 – 0,9995 = 0,0005. Por tanto, si aparecen más de quince positivos cabría pensar que muy probablemente no todos ellos sean falsos.

5) En este caso, la variable aleatoria X = "número de extranjeros en la muestra" sigue una distribución binomial de parámetros n = 1.200 y p = 0,0996. Por tanto, el valor esperado de extranjeros en la muestra es $\mu = n \cdot p$ = 119,52, es decir el promedio de extranjeros para las muestras de esas características es de, aproximadamente, 120.

Por otro lado, $P(X < 100) = P(X <= 99)$ = {Minitab o Excel} = 0,0245, es decir, es muy poco probable que una muestra contenga menos de 100 extranjeros si ésta es realmente aleatoria.

6) Se considera la variable aleatoria X = "días que dura un embarazo". Cabe tener en cuenta que $X \sim N(266,16)$.

$P(X < 240)$ = {Minitab o Excel} = 0,0521, es decir, el 5,2% de los embarazos duran menos de ocho meses.

$P(240 < X < 270) = P(X < 270) - P(X < 240)$ = {Minitab o Excel} = 0,5987 – 0,0521 = 0,5466, es decir, el 55% de los embarazos duran entre ocho y nueve meses.

Finalmente, se pide el valor c tal que $P(X > c) = 0,20$, es decir: $P(X < c) = 1 - P(X > c) = 0,80 \rightarrow$ c = {Minitab o Excel} = 279,47, es decir, el 20% de los embarazos supera los doscientos setenta y nueve días.

Inferencia de información para una población

Distribuciones muestrales y teorema central del límite. Intervalos de confianza. Contrastes de hipótesis para una población

Blanca de la Fuente

PID_00161059

UOC
Universitat Oberta de Catalunya
www.uoc.edu

Índice

Introducción

El objetivo de la inferencia estadística es obtener información acerca de una población, partiendo de la información que contiene la muestra. La selección de la muestra debe garantizar su representatividad, lo que se consigue eligiéndola al azar mediante diferentes procedimientos de muestreo que se estudian en el módulo 5.

Una vez seleccionada una muestra, se dispone de un conjunto de valores, en cuyo caso los *métodos descriptivos* estudiados en el módulo 1 facilitan el análisis de estos valores muestrales. El problema que ahora se aborda es la extensión de estos resultados al conjunto de la población o, en otras palabras, dar respuesta al siguiente interrogante: Dada cierta información muestral ¿qué podemos afirmar de la población?

La solución de este problema será el objetivo de la *inferencia estadística*.

Hasta ahora se había supuesto que los valores de los parámetros de las distribuciones de probabilidad eran conocidos. Pero esto casi nunca ocurre, de manera que tenemos que usar los datos muestrales para estimarlos. Los **estimadores** proveen valores a esos parámetros.

Cuando las inferencias que se realizan se refieren a características poblacionales concretas, es necesaria una etapa de diseño de estimadores. En este módulo se presentan los conceptos básicos para la estimación de la proporción, de la media y de la varianza de la población respectivamente.

Un enfoque alternativo es indicar un rango de valores, entre los cuales tiene que estar el parámetro con una determinada precisión: esta es la idea de un **intervalo de confianza**.

A continuación se plantea en este módulo el problema del **contraste de hipótesis**, desarrollando métodos que permiten contrastar la validez de una conjetura o de una afirmación utilizando datos muestrales. El proceso comienza cuando un investigador formula una hipótesis sobre la naturaleza de una población. La formulación de esta hipótesis implica claramente la elección entre dos opciones; a continuación, el investigador selecciona una opción basándose en los resultados de un estadístico calculado a partir de una muestra aleatoria de datos.

Objetivos

Los objetivos académicos del presente módulo se describen a continuación:

1. Explorar las distribuciones de la media, de la proporción y de la varianza muestral.

2. Aplicar el Teorema central del límite.

3. Crear intervalos de confianza.

4. Usar la distribución *t* en una prueba de hipótesis.

5. Utilizar la distribución chi-cuadrado (χ^2) en una prueba de hipótesis.

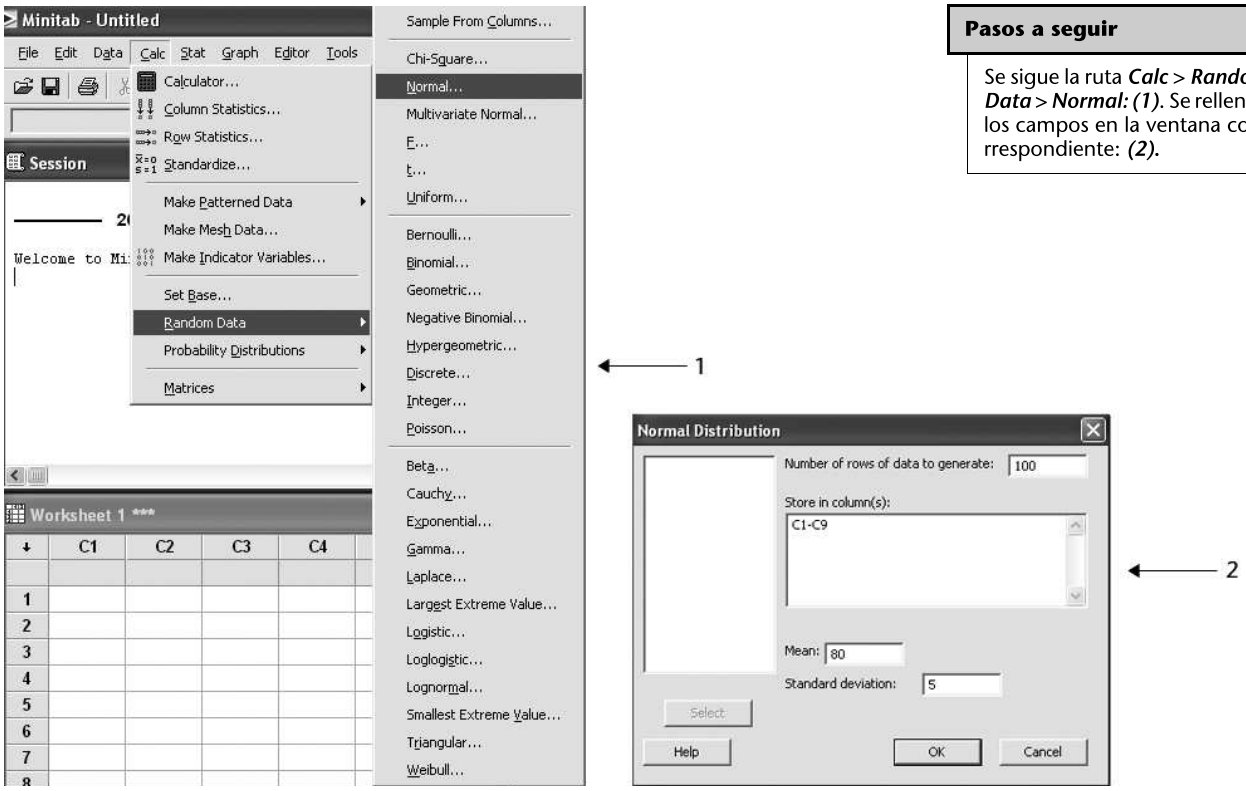
1. Distribuciones muestrales y Teorema central del límite

Una muestra aleatoria permite hacer inferencia sobre ciertas características de la distribución de la población. Esta inferencia estará basada en algún **estadístico**, es decir, alguna función particular de la información muestral. La **distribución muestral** de este estadístico es la distribución de probabilidades de los valores que puede tomar el estadístico a lo largo de todas las posibles muestras con el mismo número de observaciones, que pueden ser extraídas de la población.

Por ejemplo, en la distribución normal, los dos parámetros son la media de la población μ y la desviación estándar poblacional σ. Se puede estimar el valor μ calculando el promedio muestral o media muestral, $\bar{x}$, y el valor de σ mediante el cálculo de la desviación típica muestral, s. En este caso la media muestral, $\bar{x}$ y la desviación típica muestral, s, son los estadísticos. En el caso de la distribución binomial, los parámetros son n y p. Para estimar el parámetro proporción poblacional, p, se utiliza el estadístico proporción muestral, $\hat{p}$.

El estudio de las distribuciones muestrales se puede ilustrar mediante la creación con Minitab de 100 muestras de datos aleatorios normales con media 80 y desviación típica 5, con 9 observaciones de cada muestra (figura 1). A partir de datos aleatorios se crea una columna de datos que contenga el promedio de cada muestra o media muestral.

Figura 1. Pasos a seguir para estudiar una distribución muestral



Pasos a seguir

Se sigue la ruta ***Calc*** > ***Random Data*** > ***Normal: (1)***. Se rellenan los campos en la ventana correspondiente: ***(2)***.

Se ha generado así una matriz de nueve columnas y cien filas (figura 2). Cada componente de esta matriz es una observación aleatoria proveniente de una distribución normal de media 80 y desviación estándar 5.

Figura 2. Resultado de una matriz

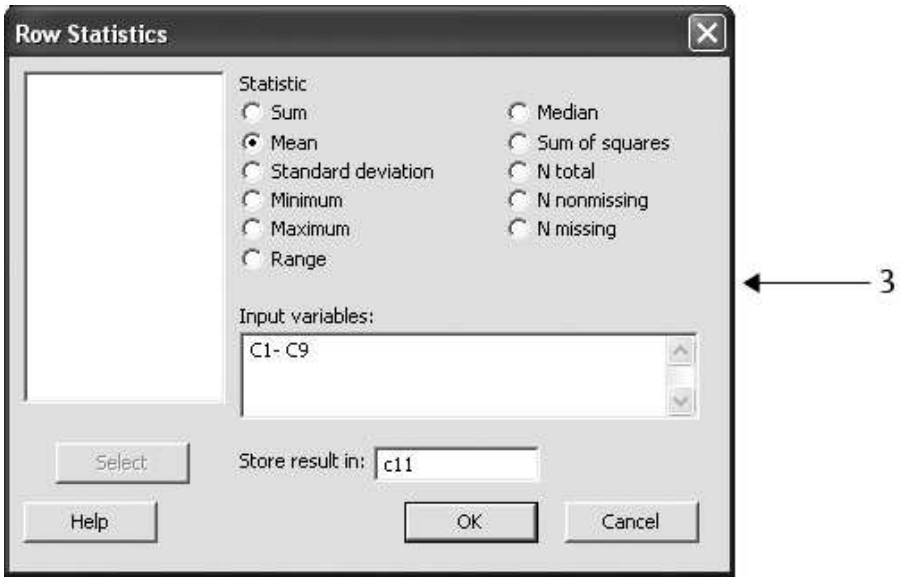
Minitab - Untitled - [Worksheet 1 ***]

File Edit Data Calc Stat Graph Editor Tools Window Help

↓	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	86,2119	79,8202	83,3841	85,2706	84,3977	73,3733	79,2915	79,7839	85,9532	
2	77,9609	77,8749	75,9217	76,5488	80,2022	77,9314	76,3676	75,2757	81,7320	
3	79,4466	81,7029	79,6951	64,8797	77,2981	83,3234	89,4473	74,5262	88,4057	
4	75,2041	79,8637	79,4250	76,8203	77,6463	74,8025	91,9961	81,1118	78,1692	
5	80,3298	80,6675	81,6815	77,6024	71,7586	84,5731	85,6013	78,1734	78,8703	
6	72,6596	75,2371	85,2919	69,2655	83,4426	89,4642	82,6151	74,9094	83,5735	
7	83,4955	79,0883	80,8709	74,2571	80,8703	83,4584	78,2343	76,5349	84,1050	
8	77,7092	76,8970	74,7272	82,4228	84,4440	78,4365	79,6917	83,6371	81,2270	
9	83,4490	81,1309	76,9926	85,5929	84,2468	78,2026	87,5854	83,7920	65,8216	
10	81,0768	85,6596	84,0062	68,5531	70,6466	76,6778	82,9853	70,8620	79,8056	
11	78,5384	77,2805	83,3829	91,9047	76,0708	73,8196	84,3317	75,8071	74,4023	
91	65,7335	73,9428	86,7681	85,7170	83,8731	84,1641	79,1929	79,6862	81,4350	
92	76,0597	83,0876	77,9049	71,7392	83,9294	90,2161	76,4049	74,5141	78,1746	
93	85,8218	83,5000	70,8513	82,4729	85,0008	78,4955	77,3336	79,6959	79,5596	
94	74,7113	68,5013	79,8273	76,5099	70,7168	86,8968	70,3262	81,5301	82,8092	
95	84,2597	70,3432	82,1207	81,9862	80,0716	79,1272	83,2129	80,0331	74,9672	
96	79,5742	78,4683	80,5376	75,5732	83,9279	80,5597	81,9438	80,2280	79,6279	
97	81,5282	81,0194	77,3914	77,9764	73,2830	76,8476	87,2748	86,6303	85,1350	
98	81,8135	82,7231	81,0083	84,1466	79,2423	81,0257	78,1474	84,3158	74,5273	
99	73,7441	79,6640	74,2215	85,9686	75,9545	81,6767	77,6596	77,1299	85,1886	
100	77,8603	80,4039	78,1897	73,7551	72,7348	75,9893	74,9260	69,2976	77,8818	

Se considera que cada una de las filas obtenidas es una muestra, y se calcula la media asociada a cada una de estas cien muestras (figura 3):

Figura 3. Pasos a seguir para calcular las medias



Pasos a seguir

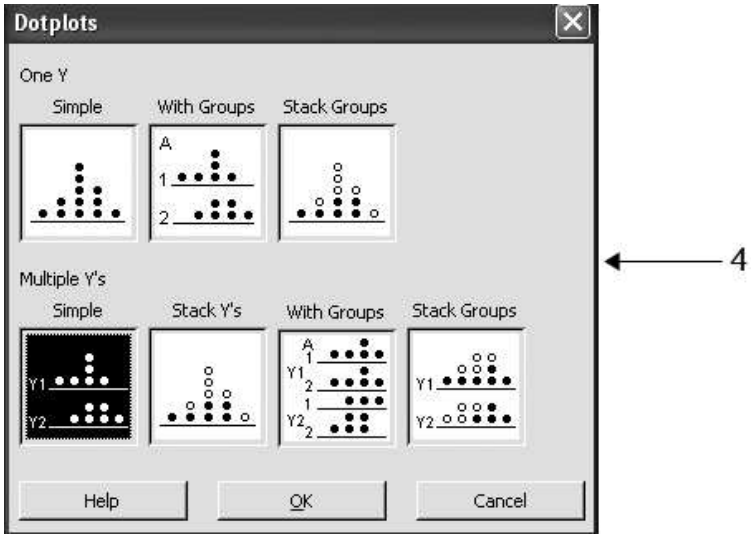
Una vez generados los datos se sigue la ruta ***Calc*** > ***Row Statistics*** y se rellenan los campos en la ventana correspondiente: ***(3)***.

En la columna C11 de la figura 4 hay cien nuevos valores (las medias). En la figura 5 se muestran los *dotplot* asociados a las columnas C1 (que representan cien valores aleatorios obtenidos de una normal 80-5) y C11:

Figura 4. Resultado del análisis

↓	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
											x-barra	
1	86,2119	79,8202	83,3841	85,2706	84,3977	73,3733	79,2915	79,7839	85,9532		81,9429	
2	77,9609	77,8749	75,9217	76,5488	80,2022	77,9314	76,3676	75,2757	81,7320		77,7572	
3	79,4466	81,7029	79,6951	64,8797	77,2981	83,3234	89,4473	74,5262	88,4057		79,8583	
4	75,2041	79,8637	79,4250	76,8203	77,6463	74,8025	91,9961	81,1118	78,1692		79,4488	
5	80,3298	80,6675	81,6815	77,6024	71,7586	84,5731	85,6013	78,1734	78,8703		79,9175	
6	72,6596	75,2371	85,2919	69,2655	83,4426	89,4642	82,6151	74,9094	83,5735		79,6065	
7	83,4955	79,0883	80,8709	74,2571	80,8703	83,4584	78,2343	76,5349	84,1050		80,1016	
8	77,7092	76,8970	74,7272	82,4228	84,4440	78,4365	79,6917	83,6371	81,2270		79,9103	
9	83,4490	81,1309	76,9926	85,5929	84,2468	78,2026	87,5854	83,7920	65,8216		80,7571	
10	81,0768	85,6596	84,0062	68,5531	70,6466	76,6778	82,9853	70,8620	79,8056		77,8081	
11	78,5384	77,2805	83,3829	91,9047	76,0708	73,8196	84,3317	75,8071	74,4023		79,5042	
12	70,0951	78,7096	77,3802	82,9569	72,4905	76,7535	88,9660	85,7643	75,4986		78,7350	

Figura 5. Pasos a seguir para crear el gráfico de puntos de los *dotplot*



Pasos a seguir

Se sigue la ruta ***Graph*** > ***Dotplot*** y se rellenan los campos en la ventana correspondiente: ***(4)***.

La salida de Minitab de la figura 6 muestra que la distribución de la variable aleatoria inicial X (columna C1) era normal y, según el gráfico de puntos, parece que también la distribución de la v.a. X-barra ($\bar{x}$) es normal, de media muy similar y desviación estándar menor (los puntos de la $\bar{x}$ están menos "dispersos" que los de la x).

También podemos hacer un histograma de frecuencias de la distribución de las medias muestrales ($\bar{x}$), como se aprecia en la figura 7.

Figura 6. Gráfico de puntos de valores de los *dotplot*

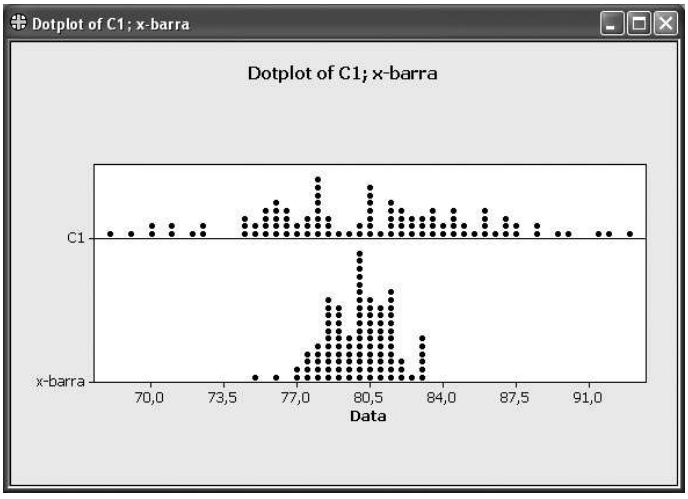
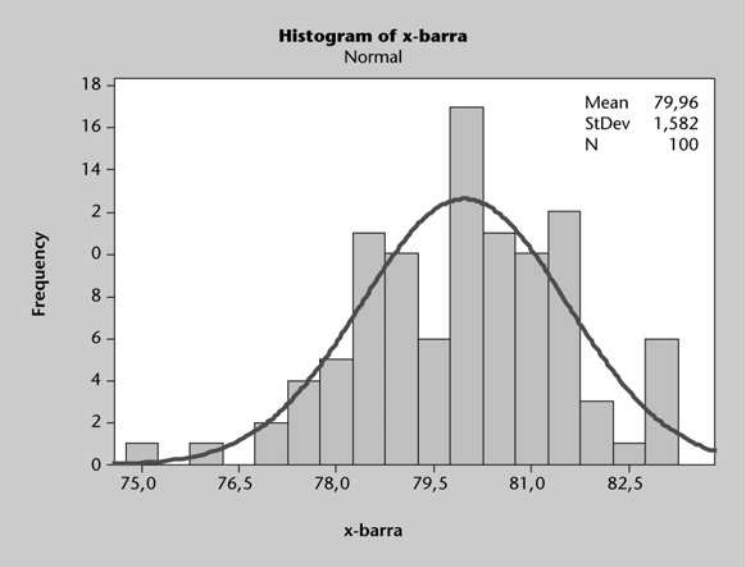


Figura 7. Histograma de frecuencias absolutas de valores de $\bar{x}$ a partir de nueve muestras aleatorias simples, cada una de tamaño cien



Finalmente, en la figura 8 se obtienen los estadísticos que describen la distribución de las medias muestrales.

Figura 8. Resultado del análisis de *X*-barra

Descriptive Statistics: x-barra

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1
Median	Q3						
x-barra	100	0	79,962	0,158	1,582	75,192	78,814
80,146	81,000						

Variable	Maximum
x-barra	83,154

Pasos a seguir

Se sigue la ruta ***Stat*** > ***Basic Statistics*** > ***Display Descriptive Statistics*** y se selecciona la variable ***C11*** (***x-barra***) en la ventana correspondiente.

La media de los cien valores contenidos de la columna C11 (y que es una aproximación a la media de la v.a. *X*-barra) es de 79,962, valor muy similar a la media de *X* (que era de 80). Esto es coherente con lo que la teoría nos indica:

- La media muestral coincide con la media de la población, $\mu_{\bar{X}} = \mu$.

La desviación estándar de los cien valores de la columna C11 (que será una aproximación a la desviación estándar de *X*-barra) es de 1,582. Si tomamos la desviación estándar de *X* (que era de 5) y la dividimos por 3 (raíz de 9, el tamaño de la muestra), obtenemos el valor 1,667.

- Ambos valores son muy parecidos, tal y como la teoría predice:

$$\sigma_{\bar{X}} = \sigma / \sqrt{n}$$

Es interesante señalar que si no se hubiera tomado inicialmente una variable normalmente distribuida, las conclusiones obtenidas serían semejantes siempre que el tamaño muestral *n* fuera lo suficientemente grande tal y como predice el **Teorema central del límite.**

Teorema central del límite

El análisis anterior se aplica sólo a la distribución normal. ¿Qué ocurre si nuestros datos provienen de otra distribución de probabilidad? ¿Podemos decir algo acerca de la distribución muestral de la media en ese caso? Para ello se utiliza el ***Teorema central del límite,*** el cual expresa que si tenemos una muestra tomada de una distribución de probabilidad con media μ y desviación típica de σ, la distribución muestral de $\bar{x}$ es aproximadamente normal con media μ y desviación típica de, $\sigma/\sqrt{n}$ que es el error estándar. Lo notable acerca del teorema central del límite es que la distribución de la media muestral de $\bar{x}$ es más o menos normal, sea cual sea la distribución original de probabilidad. A medida que aumenta el tamaño de la muestra, la aproximación a la distribución normal se acerca cada vez más.

Nota

Consideraremos que n es lo bastante grande cuando, como mínimo, $n > 30$.

Una consecuencia de este teorema es:

Dada cualquier variable aleatoria con esperanza μ y para n suficientemente grande, la distribución de la variable:

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

es una normal estándar $N(0,1)$.

Cálculo del error estándar

Recordemos que si la variable tiene una desviación típica conocida σ, el error estándar se puede calcular como $\sigma/\sqrt{n}$. Cuando σ es desconocida, calculamos el error estándar como $s/\sqrt{n}$, siendo s la desviación típica de la muestra.

Un caso particular es la **aproximación de la binomial a la normal**:

Sea X una variable aleatoria con distribución $B(n, p)$ binomial con n suficientemente grande. Entonces, X es aproximadamente normal con esperanza np y varianza $np(1–p)$.

En este caso, n grande significa que np y $np(1–p)$ son los dos mayores que 5 o bien que $n > 30$.

Por tanto, cuando el tamaño de la muestra, n, es grande, la distribución de la **proporción** es aproximadamente una distribución normal de esperanza p y desviación típica $\sqrt{p(1-p)/n}$. En este caso $\sqrt{p(1-p)/n}$, corresponde al error estándar $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$.

Recordatorio

Si X sigue una distribución **binomial** de parámetros n y p, entonces:

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

para los $k \in \{0, \ldots, n\}$

Ejemplo: se hace una encuesta sobre un determinado tema que tiene dos opciones, A y B. La probabilidad de que un individuo concreto opine A es p y n es el número de encuestas hechas. Hemos preguntado a cuatrocientos habi-

tantes y encontramos que el 30% opina A, es decir, que podemos establecer que $p = 0,3$. Entonces, la distribución de la proporción de habitantes que opina A sigue una distribución normal, cuya media es 0,3, que coincide con la proporción del 30% de los habitantes de la población que opinan A, y la desviación estándar es 0,0229, que corresponde a la desviación típica de la población dividida por la raíz cuadrada del tamaño de la muestra.

$$N\left(0,3,\sqrt{\frac{0,3(1-0,3)}{400}}\right) = N(0,3;0,0229)$$

2. Distribución de la media muestral

Se deben considerar dos casos para la distribución de la media muestral.

Caso de desviación típica poblacional conocida

Si la variable que estudiamos sigue una distribución normal con media μ y desviación típica σ conocidas, entonces la media muestral es también normal con la misma media μ y desviación típica $\sigma/\sqrt{n}$, donde n es el tamaño de la muestra.

Nota

Si σ es la desviación típica de la población y n el tamaño de la muestra, se define el **error estándar de la media muestral** como:

$$\sigma/\sqrt{n}$$

Siempre que la distribución de las medias muestrales sea una distribución normal, se puede calcular una **variable aleatoria normal estandarizada, *Z***, que tiene una media 0 y una varianza 1:

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

Observad

El error estándar es cada vez menor cuanto mayor es el tamaño de la muestra.

Si la distribución de la población no es normal pero el tamaño muestral n es suficientemente grande, entonces se usará el teorema central del límite y la variable media muestral se aproxima a una normal estándar a medida que el tamaño de la muestra aumenta. En general, dicha aproximación se considera válida para tamaños muestrales superiores a treinta.

En el apartado anterior se vio que la variable aleatoria binomial sigue una distribución normal aproximada cuando aumenta el tamaño de la muestra.

Ejemplo: en la asignatura de *Archivística* de una licenciatura de Documentación se sabe que las calificaciones siguen una distribución normal de media 7,4 y desviación estándar 0,78. Se desea conocer el porcentaje de estudiantes con nota superior a 6,5 e inferior a 8,5. ¿Con qué nota se va a calificar como "excelente" (A), si esta es la calificación del 5% de estudiantes con mejor nota?

Solución:

La variable sigue una distribución $N(7,4; 0,78)$. Primero se calcula el estadístico Z normal estandarizado:

$$P(6,5 \le X \le 8,5) = P\left(\frac{6,5-7,4}{0,78} \le \frac{X-7,4}{0,78} \le \frac{8,5-7,4}{0,78}\right) =$$
$$= P(-1,15 \le Z \le 1,41) =$$
$$= P(Z \le 1,41) - P(Z \le -1,15) = 0,9207 - 0,1251 = 0,7956$$

Los valores de probabilidad se buscan en la tabla $N(0,1)$ o calculándose con cualquier programa estadístico como se muestra en el ejemplo desarrollado en el módulo 1.

A la vista del resultado, se puede decir que el porcentaje de estudiantes con nota superior a 6,5 e inferior a 8,5 es de 79,56%.

Para calcular la nota a partir de la cual se califica como excelente, se calcula el estadístico Z normal estandarizado:

$$P(X \geq A) = P\left(\frac{X-7,4}{0,78} \geq \frac{A-7,4}{0,78}\right) = P(Z \geq z_A) = 0,05$$

En las tablas de la $N(0,1)$ o mediante cualquier programa estadístico se busca un valor z que deje a la derecha un área de 0,05, aproximadamente el valor es: $z_A = 1,645$, de manera que:

$$\frac{A-7,4}{0,78} = 1,645 \qquad \Rightarrow \qquad A = 7,4 + 1,645 \cdot 0,78 = 8,683$$

A partir de una nota de 8,6 se califica como "excelente"(A).

Caso de desviación típica poblacional desconocida

Cuando la desviación poblacional es desconocida y el tamaño de la muestra es pequeño, deberemos hacer una estimación de la desviación típica con la llamada *desviación típica muestral*. Para ello es necesario presentar una nueva distribución de probabilidad. Esta nueva distribución se conoce con el nombre de ***t* de Student** cuyas características se explicaron en el módulo 1.

Para determinar la distribución de la media muestral cuando la desviación poblacional es desconocida, se debe calcular la desviación típica muestral:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Si la variable estudiada sigue una distribución normal con media μ y desviación típica desconocida, entonces el estadístico media muestral sigue una distribución t_{n-1}, es decir, una ***t* de Student con *n*–1 grados de libertad.**

$$t_{n-1} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Los grados de libertad asociados con el valor de t son n–1(tamaño de la muestra menos uno).

Nota

En este caso se define el **error estándar de la media muestral** como:

$$s/\sqrt{n}$$

Ejemplo: el tiempo que han tardado en infectarse de virus cada uno de los ordenadores de una editorial ha sido: 2,5; 7,4; 8,0; 4,5; 7,4 y 9,2 segundos.

Suponemos que el tiempo que tarda un ordenador de esa editorial en infectarse sigue la distribución normal de media 6,5 y se desconoce la varianza poblacional. Se desea calcular la probabilidad de que un ordenador tarde entre 5 y 10 segundos en infectarse.

Solución:

Como se desconoce la varianza de la población, la media muestral seguirá una distribución ***t* de Student con 5 grados de libertad**.

Para calcular el valor del estadístico *t*, se debe calcular la desviación típica muestral. El valor obtenido es $S = 2,5$:

$$t_{n-1} = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$$

La probabilidad solicitada será:

$$p(5 \le T \le 10) = p\left(\frac{5-6,5}{2,5/\sqrt{6}} \le t_5 \le \frac{10-6,5}{2,5/\sqrt{6}}\right) = p(-1,47 \le t_5 \le 3,43) =$$
$$p(t_5 \le 3,43) - p(t_5 \le -1,47) = 0,99 - 0,1 = 0,89$$

Para calcular la probabilidad se utiliza la tabla *t* o un programa estadístico (figura 9).

Figura 9. Resultado de Minitab

```
Cumulative Distribution Function

Student's t distribution with 5 DF

   x  P( X <= x )
3,43     0,990682

Cumulative Distribution Function

Student's t distribution with 5 DF

    x  P( X <= x )
-1,47     0,100758
```

Pasos a seguir

Para calcular las probabilidades de una distribución *t* de Student se sigue la ruta ***Calc*** > ***Probability Distributions*** > ***t*** y se completan los parámetros en la ventana correspondiente. El resultado se muestra en la figura 9.

3. Distribución de la proporción muestral

En el apartado 5 del módulo 1 se dijo que la distribución binomial era la suma de n variables aleatorias independientes, cada una de las cuales tiene una probabilidad de éxito p. Para caracterizar la distribución se necesita conocer el valor de p, que es la proporción de miembros de la población que tienen una característica de interés. La **proporción muestral de éxitos** en una muestra aleatoria extraída de una población en la que la proporción de éxitos p será:

Distribución de la proporción muestral

Es una aplicación del **Teorema central del límite**.

$$\hat{p} = \frac{X}{n}$$

Por lo tanto $\hat{p}$ es la media de un conjunto de variables aleatorias independientes. Además puede utilizarse el teorema central del límite para sostener que la distribución de probabilidad de $\hat{p}$ puede considerarse una distribución normal si el tamaño de la muestra es grande.

$$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Nota

La distribución de $\hat{p}$ tiene una media igual a la proporción poblacional p.

La desviación estándar de $\hat{p}$ es el **error estándar de la media muestral** como:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Igual que en el caso de la media muestral, siempre que la distribución de la proporción muestral sea una distribución normal, se puede calcular una **variable aleatoria normal estandarizada, Z,** que tiene una media cero y una varianza uno.

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}}$$

La proporción muestral tiene muchas aplicaciones, entre las cuales se encuentra el estudio de los resultados de encuestas, la estimación de la cuota porcentual del mercado, el porcentaje de inversiones empresariales que tiene éxito y los resultados electorales entre otros.

Observad

El error estándar es cada vez menor cuanto mayor es el tamaño de la muestra.

Ejemplo: el 22% de los discos se venden por la Red en formato MP3 y el resto se vende en tiendas en formato CD. Se consideran las ventas de los próximos 5.000 discos. Se desea saber ¿qué distribución sigue la proporción muestral de discos vendidos por la Red? ¿Cuál es el número esperado de discos que se venderán por la Red? ¿Cuál es la probabilidad de que se vendan por la Red más de 1.500 discos?

Solución:

En este ejercicio se tiene que $p = 0{,}22$ y $n = 5.000$.

Para determinar la distribución de la proporción muestral, dado que el tamaño de la muestra es grande $n = 5.000$, se aplica el teorema central del límite. La distribución será aproximadamente **normal**, el valor de la media es el de la proporción poblacional (0,22).

Se calculará el error estándar $s_{\hat{p}} = \sqrt{\frac{0,22(1-0,22)}{5.000}} = 0,00586$

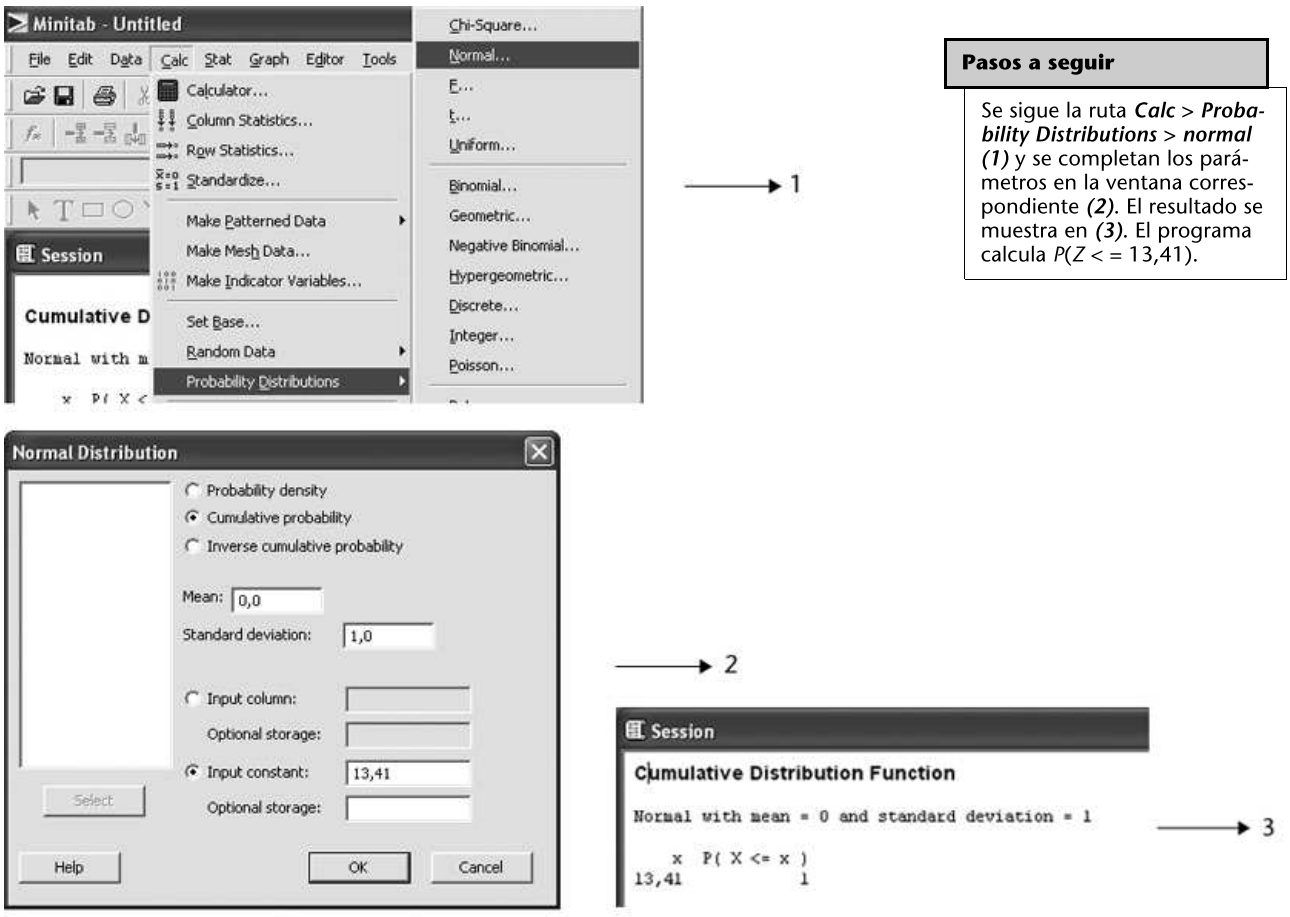
El valor esperado de discos vendidos por la Red será del 22% de los 5.000 que se venden en total, es decir, 1.100 discos en formato MP3.

La probabilidad de que se vendan menos de 1.500 discos por la Red será igual a la probabilidad de que la proporción muestral sea superior o igual al 30%. Para obtener esta probabilidad, primero se calculará el estadístico Z normal estandarizado:

$$P(p > 30\%) = P\left(Z > \frac{0,30 - 0,22}{0,00586}\right) = P(Z > 13,41) = 0$$

La probabilidad de Z se obtiene en la tabla $N(0,1)$. En la práctica, los cálculos probabilísticos anteriores se suelen automatizar con la ayuda de algún software estadístico o de análisis de datos. La figura 10 muestra cómo se pueden calcular probabilidades de una normal con ayuda de Minitab.

Figura 10. Cálculo de probabilidades con Minitab



Pasos a seguir

Se sigue la ruta ***Calc*** > ***Probability Distributions*** > ***normal*** ***(1)*** y se completan los parámetros en la ventana correspondiente ***(2)***. El resultado se muestra en ***(3)***. El programa calcula $P(Z < = 13,41)$.

El valor obtenido con Minitab es $P(Z <= 13{,}41)$. Por lo tanto, para obtener la probabilidad deseada calcularemos la probabilidad complementaria $P(Z > 13{,}41) = 1 - P(Z <= 13{,}41) = 1 - 1 = 0$.

4. Distribución de la varianza muestral

Una vez analizadas las distribuciones de las medias muestrales y las proporciones muestrales, se examinarán las distribuciones de las varianzas muestrales. A medida que las empresas y la industria ponen más énfasis en la producción de productos que satisfagan los criterios de calidad, es mayor la necesidad de calcular y reducir la varianza poblacional. Cuando la varianza es alta en un proceso, algunas características de los productos pueden tener una gama más alta de valores, como consecuencia de la cual hay más productos que no tienen un nivel de calidad aceptable. Se pueden obtener productos de calidad si el proceso de producción tiene una varianza baja, de manera que es menor el número de unidades que tienen un nivel de calidad inferior al deseado. Comprendiendo la distribución de las varianzas muestrales podemos hacer inferencias sobre la varianza poblacional.

Si se estudia una muestra aleatoria de tamaño n y varianza muestral s^2 obtenida de una población normal de media μ y varianza σ^2 desconocidas, entonces la varianza muestral se distribuye como una χ^2_{n-1} con n–1 grados de libertad:

$$\chi^2_{n-1} = \frac{(n-1)s_x^2}{\sigma_x^2}$$

Por lo tanto, se pueden hacer inferencias sobre la varianza poblacional σ^2 utilizando s^2 y la distribución chi-cuadrado. Este proceso se muestra en el siguiente ejemplo.

Ejemplo: en una gran ciudad se ha observado que durante el verano las facturas del consumo de electricidad siguen una distribución normal que tiene una desviación típica del 100 euros. Se ha tomado una muestra aleatoria de 25 facturas. Se desea calcular la probabilidad de que la desviación típica muestral sea inferior a 75 euros.

Solución:

En este ejercicio se tiene que $n = 25$ y $\sigma^2 = (100)^2$. Utilizando la distribución chi-cuadrado se puede establecer que:

$$P\left(s^2 < 75^2\right) = P\left(\frac{(25-1)75^2}{(100)^2} < \chi^2_{24\,g.l.}\right) = P\left(13,5 < \chi^2_{24\,g.l.}\right)$$

Los valores de la distribución chi-cuadrado pueden obtenerse en la tabla de dicha distribución con 24 grados de libertad:

$$\chi^2_{24\,g.l.} = 12,401;\ \chi^2_{24\,g.l.} = 13,848$$

El valor de probabilidad estará entre 0,025 y 0,05 (0,0428) exactamente.

5. Intervalos de confianza para una población

En los apartados anteriores hemos considerado la estimación puntual de un parámetro desconocido de la población, es decir, el cálculo de un único número que sea una buena aproximación. En la mayoría de los problemas prácticos, un estimador puntual por sí solo es inadecuado. Por ejemplo, supongamos que un control hecho sobre una muestra aleatoria de manuales procedentes de un gran envío de una editorial nos lleva a estimar que el 10% de todos los manuales son defectuosos. Un gerente que se enfrenta a este dato posiblemente se hará preguntas del tipo: ¿puede estar totalmente seguro de que el verdadero valor del porcentaje de manuales defectuosos está entre el 5% y el 15%? O ¿es muy posible que entre el 9% y el 11% de los manuales sean defectuosos? Esta clase de preguntas requieren información que va más allá de la contenida en una simple estimación puntual; son preguntas que buscan la fiabilidad de dicho estimador. En otras palabras, se trata de la búsqueda de un **estimador por intervalos**, un rango de valores entre los que posiblemente se encuentre la cantidad que se estima.

Debemos medir de alguna manera la confianza que podemos tener en el intervalo. Este porcentaje de muestras que dan lugar a intervalos que contienen el auténtico valor del parámetro es el llamado **nivel de confianza**.

Nivel de confianza

El nivel de confianza también se denota por $(1 - \alpha)$ 100% normalmente consideraremos $(1 - \alpha)$, igual a 90%, 95% o 99%.

Así pues, un intervalo de confianza para cierto parámetro con un nivel de confianza de C% es un intervalo calculado a partir de una muestra de manera que el procedimiento de cálculo garantiza que el C% de las muestras dé lugar a un intervalo que contenga el valor real del parámetro.

La expresión *confianza del 95%* indica confianza en el método utilizado, de manera que el 95% de las veces que apliquemos el método a la misma población obtendremos intervalos que sí contienen el valor del parámetro poblacional.

Intervalo de confianza para la media cuando la población es normal y conocemos la desviación estándar

La variable que queremos estudiar sigue una ley normal de media μ (desconocida) y desviación estándar σ conocida. Disponemos de una muestra aleatoria simple de tamaño n y el valor de la media de la muestra es $\overline{x}$.

Se calculan los intervalos de confianza al nivel de confianza $(1 - \alpha)$% mediante la siguiente expresión:

$$(\text{media de la muestra} - ME, \text{media de la muestra} + ME)$$

donde ME es el **margen de error** que tenemos que calcular, de manera que el $(1 - \alpha)$ % de las muestras produzca un intervalo que contenga el verdadero valor de μ.

El procedimiento que describimos sirve también para variables que no sigan una distribución normal, siempre que la desviación típica sea conocida y que el tamaño de la muestra sea $n > 30$.

Fijamos el nivel de confianza: se acostumbra a considerar $(1 - \alpha)$ igual a 90%, 95% o 99%.

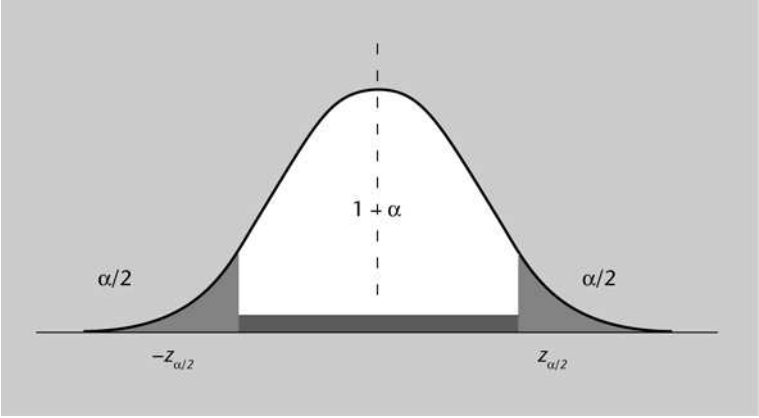
Calculamos el **error estándar** de la media como $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

Obtenemos el **valor crítico**, que es aquel valor $z_{\alpha/2}$ que hace que:

$$P(Z \geq z_{\alpha/2}) = \alpha/2$$

en el que Z es una variable aleatoria normal $N(0,1)$. Se muestra gráficamente en la figura 11.

Figura 11. Gráfico de intervalo de confianza para μ con desviación típica conocida



Para los niveles de confianza usuales, los valores críticos correspondientes son:

- $(1 - \alpha) = 90\% = 0{,}9$, $\alpha = 0{,}1$ y $z_{\alpha/2} = z_{0,05} = 1{,}645$
- $(1 - \alpha) = 95\% = 0{,}95$, $\alpha = 0{,}05$ y $z_{\alpha/2} = z_{0,025} = 1{,}96$
- $(1 - \alpha) = 99\% = 0{,}99$, $\alpha = 0{,}01$ y $z_{\alpha/2} = z_{0,005} = 2{,}575$

Calculamos el denominado ***margen de error*** (también denominado ***precisión de la estimación***) como $z_{\alpha/2}$ para el error estándar, es decir, como:

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Nota

Por tanto, el margen de error es la mitad de la longitud del intervalo de confianza.

El intervalo de confianza obtenido con la muestra de partida es:

$$\left(\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) = \left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

o lo que es lo mismo, $\bar{x} \pm ME$.

Es necesario interpretar exactamente los intervalos de confianza. Si se extraen repetida e independientemente muestras aleatorias de n observaciones de la población, entonces el $100(1-\alpha)\%$ de estos intervalos contendrá el verdadero valor de la media poblacional.

El efecto del tamaño de la muestra

En muchas ocasiones, una vez fijado el nivel de confianza nos marcaremos como objetivo dar el valor del parámetro μ con cierta precisión. La única manera de obtener la precisión deseada consiste en modificar de forma adecuada el tamaño de la muestra. Supongamos que deseamos una precisión o margen de error ME; puesto que sabemos que:

Tamaño de la muestra

Es fácil ver que si queremos reducir el ancho del intervalo de confianza a la mitad, deberemos tomar una muestra cuatro veces mayor.

$$ME = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Se obtiene el tamaño deseado de la muestra para dicha precisión:

$$n \geq \left(z_{\alpha/2} \right)^2 \frac{\sigma^2}{ME^2}$$

Intervalo de confianza para la media cuando la población es normal y desconocemos la desviación estándar

La variable que queremos estudiar sigue una ley normal de media μ (desconocida) y desviación estándar también desconocida. Disponemos de una muestra aleatoria simple de tamaño n y el valor de la media de la muestra es $\bar{x}$. Entonces:

Calculamos los intervalos de confianza al nivel de confianza $(1-\alpha)\%$, mediante la siguiente expresión se fija el **nivel de confianza**, que habitualmente se escribe como $(1-\alpha)\%$.

Calculamos la desviación típica muestral S para obtener el **error estándar** de la media como:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Calculamos el **valor crítico**, que es aquel valor $t_{\alpha/2}$ tal que:

$$P(t_{n-1} \geq t_{n-1,\alpha/2}) = \alpha/2$$

en el que t_{n-1} es una variable aleatoria de Student con $n-1$ grados de libertad.

Como el **margen de error** es:

$$ME = t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

El intervalo de confianza obtenido con la muestra es el siguiente:

$$\overline{x} \pm ME$$

Intervalo de confianza para la proporción

Interesa conocer la proporción de miembros de la población que poseen una característica específica. Si se toma una muestra aleatoria simple de tamaño *n,* la proporción muestral es un buen estimador de la proporción poblacional. En este apartado se desarrollan intervalos de confianza para la proporción.

Cuando el tamaño de la muestra sea bastante grande, en concreto siempre que el tamaño sea superior a cien, se aplicará el teorema centra del límite, y, como se ha visto en apartados anteriores, la distribución de la proporción muestral sigue una distribución normal estándar $N(0,1)$.

Igual que en los intervalos anteriores se calcula el ***margen de error*** como $z_{\alpha/2}$ multiplicado por el error estándar, es decir:

$$ME = z_{\alpha/2}s_{\hat{p}} = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Nota

El parámetro es *p*.

El estadístico es $\hat{p}$.

El intervalo de confianza obtenido con la muestra de partida será:

$$\hat{p} \pm ME$$

El tamaño de la muestra es $n = \left(z_{\alpha/2}\right)^2 \frac{\hat{p}(1-\hat{p})}{ME^2}$

Ejemplo: un servidor de correo ha recibido 2.000 mensajes, de los cuales 250 son "SPAM". Construid un intervalo de confianza del 96% para la proporción de mensajes "SPAM", ¿cuántos correos se han de estudiar en el servidor para poder afirmar que el error entre la proporción de mensajes "SPAM" recibidos y la probabilidad de que el servidor reciba un "SPAM" sea menor que 0,03 con una probabilidad del 95%?

Solución:

El intervalo de confianza del 96% para la proporción de la población se obtiene por medio de la ecuación:

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}\right)$$

Se deduce que $\hat{p} = \frac{250}{2000} = 0{,}125,\ n = 2000,\ z_{\alpha/2} = z_{0{,}02} = 2{,}054.$

Por lo tanto, el intervalo de confianza de la proporción poblacional al 96% es

$$\left(0{,}125 - 2{,}054\sqrt{\frac{0{,}125\cdot 0{,}875}{2000};}\quad 0{,}125 + 2{,}054\sqrt{\frac{0{,}125\cdot 0{,}875}{2000}}\right) = (0{,}1098;\ 0{,}1402).$$

Se podría decir que la proporción de todos los mensajes Spam recibidos de la población estarán entre el 10,98% y el 14,02% (con un margen de error del 1,52% al nivel de confianza del 96%).

Se calculará el mínimo tamaño de la muestra necesario para que el error sea menor que 0,03 con una probabilidad del 95% es:

$$n \geq (z_{\alpha/2})^2 \frac{\hat{p}\cdot(1-\hat{p})}{ME^2} = (z_{0{,}025})^2 \frac{0{,}125\cdot 0{,}875}{0{,}03^2} = 1{,}96^2 \cdot \frac{0{,}109}{0{,}0009} = 466{,}75$$

Por tanto, se deben estudiar 467 mensajes.

Ejemplo con Minitab: en el ejemplo anterior se comparan los intervalos de confianza al 90 y el 99%, manteniendo constante el tamaño de la muestra, para contestar a la siguiente pregunta: Conforme aumenta la amplitud de un intervalo de confianza, ¿aumenta o disminuye el nivel de confianza asociado? En las figuras 12 y 13 utilizamos Minitab para analizar ambos escenarios.

Figura 12. Resultado del Intervalo de confianza del 90% con Minitab

Test and CI for One Proportion

Test of p = 0,125 vs p not = 0,125

Sample	X	N	Sample p	90% CI	Z-Value	P-Value
1	250	2000	0,125000	(0,112836; 0,137164)	0,00	1,000

Using the normal approximation.

Figura 13. Resultado del Intervalo de confianza del 99% con Minitab

Test and CI for One Proportion

Test of p = 0,125 vs p not = 0,125

Sample	X	N	Sample p	99% CI	Z-Value	P-Value
1	250	2000	0,125000	(0,105951; 0,144049)	0,00	1,000

Using the normal approximation.

Notar que al aumentar el nivel de confianza, deberemos ampliar la amplitud del intervalo a fin de "abarcar" un rango mayor para el parámetro poblacional estimado.

Intervalo de confianza para la varianza

¿Cómo se puede construir un intervalo de confianza para la varianza poblacional?

Primero se fijará el nivel de confianza $1 - \alpha$. Se calculará el estadístico.

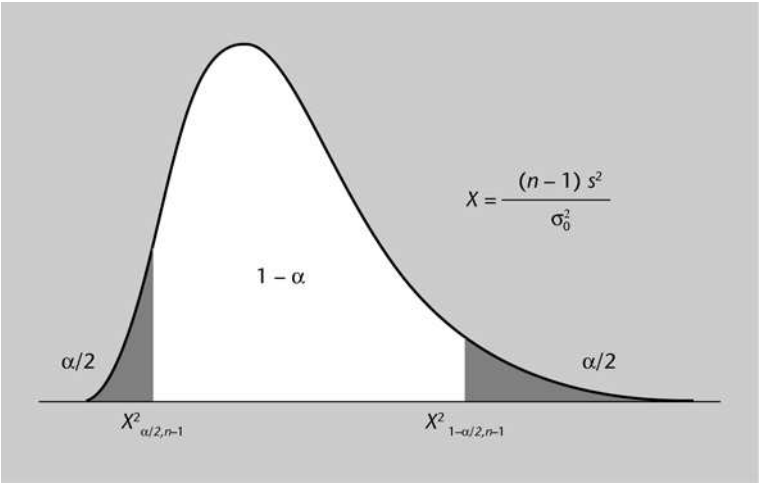
$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

es una observación de una variable aleatoria χ^2 con $n - 1$ grados de libertad.

Donde s^2 es la varianza muestral de una muestra aleatoria de tamaño n tomada de una población normal de varianza σ^2.

La figura 14 muestra los valores de la distribución χ^2_{n-1} que cortan una probabilidad de $\alpha/2$ en las dos colas, es decir, los puntos críticos $\chi^2_{n-1,\alpha/2}$ y $\chi^2_{n-1,\,1-\alpha/2}$.

Figura 14. Gráfico de intervalo de confianza de la varianza



Ejemplo de intervalo de confianza para la varianza

Una empresa de autobuses urbanos espera que las horas de llegada en diversas paradas tengan poca variabilidad. La varianza de la muestra de 10 tiempos de llegada de autobús fue $s^2 = 4{,}8$ minutos^2. Suponiendo que la población de tiempos de llegada tiene una distribución normal, se desea determinar un intervalo de confianza del 95% para la varianza poblacional de los tiempos de llegada.

El estadístico de prueba: $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$ tiene una distribución chi-cuadrado con $n - 1 = 9$ grados de libertad. Determinamos los valores $\chi^2_{9,0,975} = 16{,}0471$ y $\chi^2_{9,0,025} = 45{,}7222$.

El intervalo de confianza para la varianza de la población será:

$$\left[\frac{(n-1)s^2}{\chi^2_{n-1,\,1-\alpha/2}};\frac{(n-1)s^2}{\chi^2_{n-1,\,\alpha/2}}\right]=\left[\frac{9\cdot 4{,}8}{45{,}7222};\frac{9\cdot 4{,}8}{16{,}0471}\right]=[0{,}94;2{,}69]\text{minutos}$$

La raíz cuadrada de esos valores será el intervalo de confianza de 95% para la desviación estándar: $0{,}97 \leq \sigma \leq 1{,}64$.

6. Contrastes de hipótesis para una población

En este apartado se desarrollan métodos para contrastar hipótesis que permiten comparar la validez de una conjetura o afirmación utilizando datos muestrales. El proceso comienza cuando un investigador formula una hipótesis sobre la naturaleza de una población. La formulación de esta hipótesis implica la elección entre dos opciones; a continuación, el investigador selecciona una opción basándose en los resultados de un estadístico calculado a partir de una muestra aleatoria de datos.

He aquí algunos ejemplos de problemas representativos:

1) Un investigador quiere saber si una propuesta de reforma fiscal es acogida de igual forma por hombres y mujeres. Para analizar si es así, recoge las opiniones de una muestra aleatoria de hombres y mujeres.

2) Una compañía recibe un cargamento de piezas. Sólo puede aceptar el envío si no hay más de un 5% de piezas defectuosas. La decisión de si aceptar la remesa puede basarse en el examen de una muestra aleatoria de piezas.

3) Una profesora está interesada en valorar la utilidad de hacer controles regularmente en un curso de estadística. El curso consta de dos partes y la profesora realiza estos controles sólo en una de ellas. Cuando acaba el curso, compara los conocimientos de los estudiantes en las dos partes del curso mediante un examen final y analiza la hipótesis de que los controles aumentan el nivel medio de conocimientos.

Los ejemplos propuestos tienen algo en común. La hipótesis se formula sobre la población y las conclusiones sobre la validez de esta hipótesis se basan en la información muestral. El test o contraste será la herramienta que nos permitirá extraer conclusiones a partir de la diferencia entre las observaciones y los resultados que se deberían obtener si la hipótesis de partida fuese cierta.

Planteamiento del contraste de hipótesis

Hipótesis

Con la misma hipótesis nula podemos estudiar varias hipótesis alternativas.

En la prueba de hipótesis se comienza proponiendo una hipótesis de partida acerca de un parámetro poblacional. Esta hipótesis se llama **hipótesis nula** y se representa como $\boldsymbol{H_0}$. A continuación se define otra hipótesis, la **hipótesis alternativa**, que es la opuesta de lo que se afirma en la hipótesis nula. La hipótesis alternativa se representa como $\boldsymbol{H_1}$. El procedimiento para probar una hipótesis comprende el uso de datos de una muestra para probar las dos aseveraciones representadas por H_0 y H_1.

Las hipótesis expresan una afirmación sobre el valor del parámetro. Podemos tener una hipótesis nula del tipo $\boldsymbol{H_0}$: $\theta = \theta_0$.

La hipótesis alternativa puede ser unilateral, como $\boldsymbol{H_1}$: $\theta > \theta_0$ o $\boldsymbol{H_1}$: $\theta < \theta_0$, o bilateral, como $\boldsymbol{H_1}$: $\theta \neq \theta_0$.

Una vez planteadas las hipótesis nula y alternativa, debemos tomar una decisión a partir de las observaciones. Por otro lado, existen dos decisiones posibles:

1) Aceptar la hipótesis nula.

2) Rechazar la hipótesis nula.

Errores en el contraste

Con el fin de llegar a una de estas dos conclusiones, se adopta una **regla de decisión** basada en la evidencia muestral. Por consiguiente, *no se puede saber con seguridad* si la hipótesis nula es cierta o falsa. Por tanto, cualquier regla de decisión adoptada tiene cierta probabilidad de llegar a una conclusión falsa. Como se indica en la tabla 1, pueden cometerse dos tipos de errores. Un error que se puede cometer, llamado **error de tipo I**, es rechazar una hipótesis nula cierta. Si la regla de decisión es tal que la probabilidad de rechazar la hipótesis nula cuando es cierta es α, entonces α se llama **nivel de significación** del contraste. La probabilidad de aceptar la hipótesis nula cuando es cierta es $(1 - \alpha)$. El otro error posible, llamado **error de tipo II**, ocurre cuando se acepta una hipótesis nula falsa. La probabilidad de cometer este tipo de error, cuando la hipótesis nula es falsa, se denota por β. Entonces, la probabilidad de rechazar una hipótesis nula falsa es $(1 - \beta)$, y se denomina **potencia del contraste.**

Regla de decisión

Error de tipo I: rechazar una hipótesis nula cierta.

Error de tipo II: aceptar una hipótesis nula falsa.

Nivel de significación: la probabilidad de rechazar una hipótesis nula que es cierta (esta probabilidad a veces se expresa en %, con lo que nos referimos a un contraste de significación α como un contraste al nivel 100 α%).

Potencia: la probabilidad de rechazar una hipótesis nula que es falsa.

Tabla 1. Errores y decisiones correctas en contrastes de hipótesis

		Condición de la población	
		H_0 **verdadera**	H_0 **falsa**
Decisión	**Aceptar H_0**	Decisión correcta	Error de tipo II
	Rechazar H_0	Error de tipo I	Decisión correcta

Atención

Un nivel α = **0,05** significa que aunque la hipótesis nula sea cierta, los datos de cinco de cada cien muestras nos la harán rechazar. Es decir, aceptamos que podemos rechazar la hipótesis nula equivocadamente cinco de cada cien veces.

Para plantear y resolver un contraste de hipótesis, es necesario:

1) Fijar las hipótesis nula y alternativa.

2) Fijar un nivel de significación.

3) Determinar el estadístico de contraste y su ley.

4) A partir de aquí, tenemos dos métodos posibles:

4a) Calcular el *p*-valor asociado a nuestro estadístico de contraste calculado. Comparar el *p*-valor con el nivel de significación y tomar una decisión.

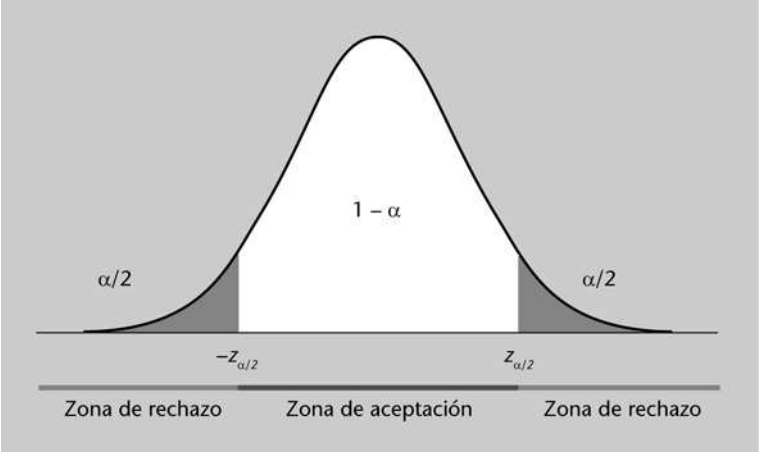
4b) Calcular el valor crítico. Comparar el valor crítico con el estadístico de contraste y tomar una decisión.

Zona de aceptación y zona de rechazo de la hipótesis nula

Ejemplo 1. "Contraste bilateral"

La parte del gráfico (figura 15) sombreada en rojo corresponde a la zona en la que rechazamos la hipótesis nula. La zona sin sombrear corresponde a la región de aceptación de la hipótesis nula.

Figura 15. Gráfico que muestra la zona de aceptación y de rechazo de la hipótesis nula en un contraste bilateral



Recordad

Si tenemos una muestra de tamaño n de una distribución $N(\mu, \sigma^2)$, entonces

$$Z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

sigue una distribución normal estándar.

Para determinar el valor $z_{\alpha/2}$, sólo hay que imponer que el error de tipo I (probabilidad de rechazar H_0 cuando es cierta) sea menor o igual que el nivel de significación α. Por ejemplo, para $\alpha = 0{,}05$ encontramos (por ejemplo, en las tablas de la normal) que $z_{\alpha/2} = 1{,}96$.

Para decidir si rechazamos la hipótesis nula o no, usaremos el llamado **estadístico de contraste.** Un estadístico de contraste es una función de la muestra cuya distribución conocemos bajo la hipótesis nula.

Validez del método

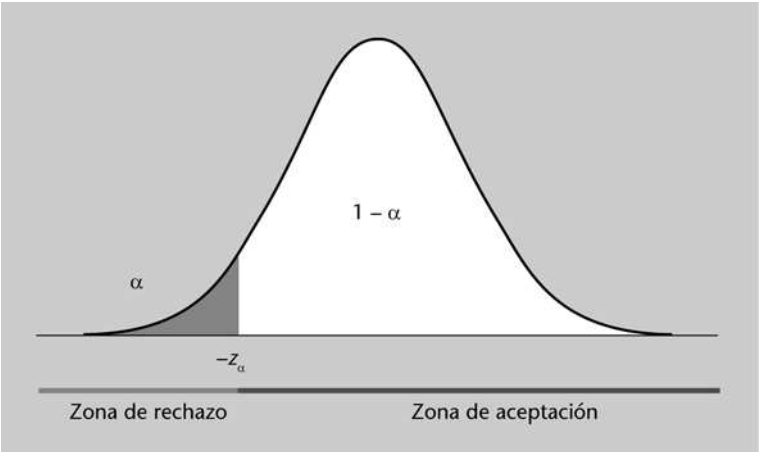
El método es el mismo para cualquier distribución simétrica, así que también sirve si el estadístico de contraste sigue una distribución t de Student.

- Aceptaremos H_0 si $|z| \leq z_{\alpha/2}$
- Rechazaremos H_0 si $|z| \geq z_{\alpha/2}$

Ejemplo 2. "Contraste unilateral inferior"

La parte del gráfico (figura 16) sombreada corresponde a la zona de rechazo de la hipótesis nula. La zona sin sombrear corresponde a la región de aceptación de la hipótesis nula.

Figura 16. Gráfico que muestra la zona de rechazo de la hipótesis nula en un contraste unilateral inferior



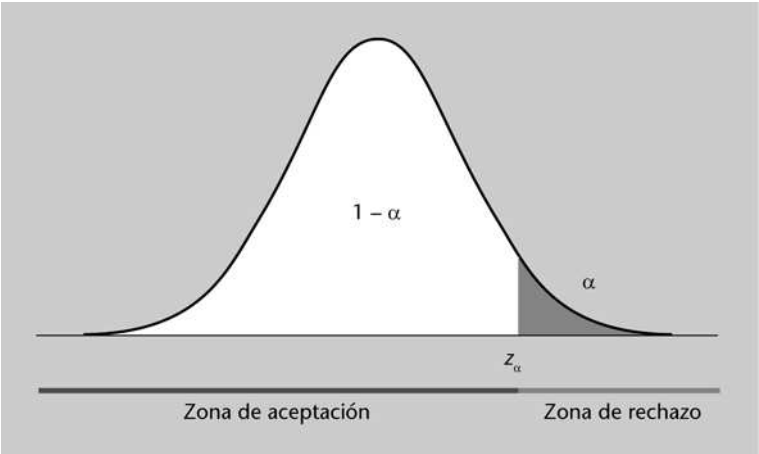
Para $\alpha = 0{,}05$ encontramos que $-z_\alpha = -1{,}65$. En este contraste unilateral se dice que la probabilidad de la cola de la izquierda debe ser α.

- Aceptaremos H_0 si $\boldsymbol{Z} \geq -z_\alpha$
- Rechazaremos H_0 si $\boldsymbol{Z} < -z_\alpha$

Ejemplo 3. "Contraste unilateral superior"

La parte del gráfico (figura 17) sombreada en rojo corresponde a la zona en la que rechazamos la hipótesis nula. La zona sin sombrear corresponde a la región de aceptación de la hipótesis nula.

Figura 17. Gráfico que muestra la aceptación o no de la hipótesis nula en un contraste unilateral superior



Para $\alpha = 0{,}05$ encontramos que $z_\alpha = 1{,}65$. En este contraste unilateral se dice que la probabilidad de la cola de la derecha debe ser α.

- Aceptaremos H_0 si $\mathbf{Z} \leq z_\alpha$
- Rechazaremos H_0 si $\mathbf{Z} > z_\alpha$

El *p*-valor

Existe otro método para examinar el contraste de la hipótesis nula. Obsérvese que si se utiliza un nivel de significación bajo se reduce la probabilidad de rechazar una hipótesis nula verdadera. Eso modificaría la regla de decisión para que fuera menos probable que se rechazara la hipótesis nula, independientemente de que fuera verdadera o no. Evidentemente, cuanto menor es el nivel de significación al que se rechaza una hipótesis nula mayores son las dudas sobre su veracidad. En lugar de contrastar hipótesis a los niveles preasignados de significación, los investigadores a menudo hallan el nivel menor de significación al que se puede rechazar una hipótesis nula.

El *p*-valor es el menor nivel de significación al que puede rechazarse una hipótesis nula.

El criterio del *p*-valor es: rechazar H_0 si el *p*-valor $< \alpha$.

Interpretación del *p*-valor

Se considera una muestra aleatoria de n observaciones procedentes de una población que sigue una distribución normal de media μ y desviación estándar σ y la media muestral calculada $\bar{x}$. Se ha contrastado la hipótesis nula $H_0 : \mu = \mu_0$ frente a la alternativa $H_1 : \mu > \mu_0$

El *p*-valor del contraste es:

$$p\text{-valor} = P\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_p \,|\, H_0 : \mu = \mu_0 \right)$$

donde z_p es el valor normal estándar correspondiente al menor valor de significación al que puede rechazarse la hipótesis nula. La mayoría de los programas informáticos estadísticos calculan el *p*-valor, este suministra más información sobre el contraste basándose en la media muestral observada, por lo que se utiliza frecuentemente en muchas aplicaciones estadísticas.

Ejemplo de aplicación del *p*-valor: un grupo editorial emite un periódico especializado en información económica. El director del periódico desea saber si el número medio de ejemplares diarios producidos y no vendidos es menor de 400. Para dar respuesta a esta pregunta, se toma una muestra formada por los resultados correspondientes a 172 días elegidos de forma aleatoria. La media de dicha muestra es de 407 ejemplares no vendidos, con una desviación estándar de 38.

Utilizando un nivel de significación de 0,05, realizad un contraste de hipótesis para responder razonadamente a la pregunta del director del periódico.

Solución:

1) Si se hace el contraste H_0: media poblacional = 400 contra H_1: media poblacional $\neq$ 400.

Primero se calcula el estadístico de contraste para decidir si rechazamos la hipótesis nula o no.

La desviación estándar de la muestra es: $\frac{S}{\sqrt{n}} = \frac{38}{\sqrt{172}} = 2,89.$

El estadístico será $z = \frac{407 - 400}{2,89} = 2,42$, este valor es una observación de una distribución $N(0,1)$.

En este caso, por ser un contraste bilateral se divide el nivel de significación α por igual entre las dos colas de la distribución normal. Por lo tanto, la probabilidad de que Z sea superior $z_{\alpha/2}$ o inferior a $-z_{\alpha/2}$ es α. En este caso, el

p-valor es la suma de las probabilidades de la cola superior y la cola inferior. El *p*-valor correspondiente al contraste de dos colas es:

$$p - valor = 2P\left(\left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right| > z_{\alpha/2}\right);$$

$$P\left(Z > |2,42|\right) = P\left(Z > 2,42\right) + P\left(Z < -2,42\right) = 2 \cdot 0,00776 = 0,01552$$

Como 0,01552 es menor que el nivel de significación propuesto ($\alpha = 0,05$), se rechaza la hipótesis nula. No se puede afirmar que el número medio de ejemplares diarios producidos y no vendidos sea de 400. Se acepta que es distinto de 400.

2) Si se hace el contraste H_0: media poblacional = 400 contra H_1: media poblacional > 400, entonces el *p*-valor es la probabilidad "es la cola de la derecha":

$$p - valor = P\left(Z\right) > z_{\alpha}$$

$P\left(Z > 2,42\right) = 0,00776 < \alpha \Rightarrow$ Se rechaza la hipótesis nula.

Se acepta la hipótesis alternativa, por lo tanto, se acepta que el número medio de ejemplares diarios producidos y no vendidos es mayor de 400.

3) Si se realiza el contraste H_0: media poblacional = 400 contra H_1: media poblacional < 400, entonces el p-valor es la probabilidad "es la cola de la izquierda":

$$p - valor = P\left(Z\right) < z_{\alpha}$$

$P\left(Z < 2,42\right) = 1 - 0,00776 = 0,99224 > \alpha \Rightarrow$ No se puede rechazar la hipótesis nula.

Se rechazará la hipótesis alternativa, luego el número medio de ejemplares diarios producidos y no vendidos no es menor de 400.

Por tanto, a la vista de los resultados de los tres contrastes, la contestación a la pregunta del director sería:

"El número medio de ejemplares diarios producidos y no vendidos es mayor de 400".

Para calcular el *p*-valor se suele utilizar un software estadístico, como se verá en ejemplos resueltos con Minitab.

Otro procedimiento: para resolver contrastes bilaterales utilizando intervalos de confianza.

Ejemplo: supongamos que se plantea el siguiente contraste bilateral:

$$H_0: \mu = 280,\ H_1: \mu \neq 280$$

Para probar esta hipótesis con un nivel de significación $\alpha = 0{,}05$, el tamaño de la muestra es 36 y se determinó que la media muestral $\bar{x} = 278{,}5$ y la desviación estándar de las muestras $s = 12$. Sustituyendo estos resultados con $z_{0{,}025} = 1{,}96$, vemos que el intervalo de confianza del 95% para la media de la población es:

$$\bar{x} \pm 1{,}96\frac{s}{\sqrt{n}} \;;\; 278{,}5x \pm 1{,}96\frac{12}{\sqrt{36}} \;;\; 278{,}5 \pm 3{,}92$$

El intervalo será: (274,58; 282,42).

El resultado permite llegar a la conclusión de que, con un 95% de confianza, la media para la población está entre 274,58 y 282,42. Como el valor supuesto de la media de la población $\mu_0 = 280$ está en el intervalo de confianza, la conclusión del contraste es que no se puede rechazar la hipótesis nula, por tanto, aceptamos la hipótesis de que: H_0: $\mu = 280$.

Ejemplo de inferencia para una población (utilizando Minitab)

Una característica importante en el diseño de una página web es el tiempo que el usuario tardará en abrir la página, que se considera una variable normal. Con el objetivo de estimar el tiempo medio, se seleccionan al azar 101 páginas, entre las que ha diseñado una empresa el último año, obteniéndose los datos siguientes (en centésimas de segundo):

Tabla 2. Tiempo de descarga de páginas web

Tiempo de descarga	55	60	62	64	65	69
Número de páginas	11	21	26	19	15	9

Observación: se crea un fichero de datos en la hoja de Minitab, introduciendo los datos de forma unitaria.

a) Se comprueba que la colección de datos sigue una distribución aproximadamente normal.

b) Puede considerarse que el tiempo medio de apertura de las páginas de esta empresa es de 62 centésimas de segundo, con un nivel de confianza del 90%. ¿Qué resultado se obtiene? Razónese la respuesta del contraste a través del *p*-valor.

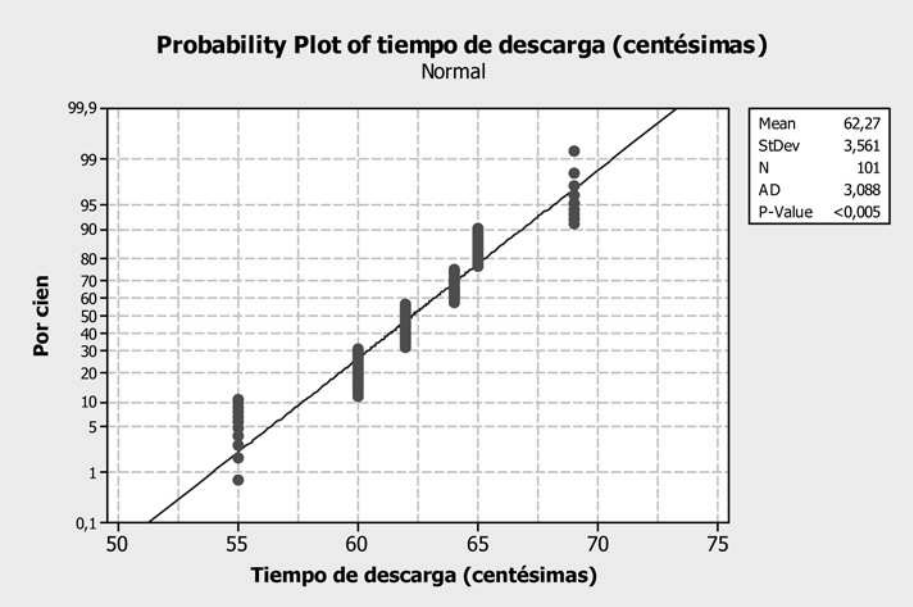
c) Calcúlese un intervalo de confianza a nivel del 90% para el tiempo medio y coméntese si el resultado obtenido es coherente con el resultado esperado.

d) Finalmente, se realizará el mismo contraste que en el apartado b), pero suponiendo esta vez que no se conoce la desviación estándar.

Solución:

a) Para comprobar la normalidad de los datos, se selecciona ***Stat*** > ***Basic Statistics*** > ***Normality Test***. Así se obtiene el gráfico de la figura 18.

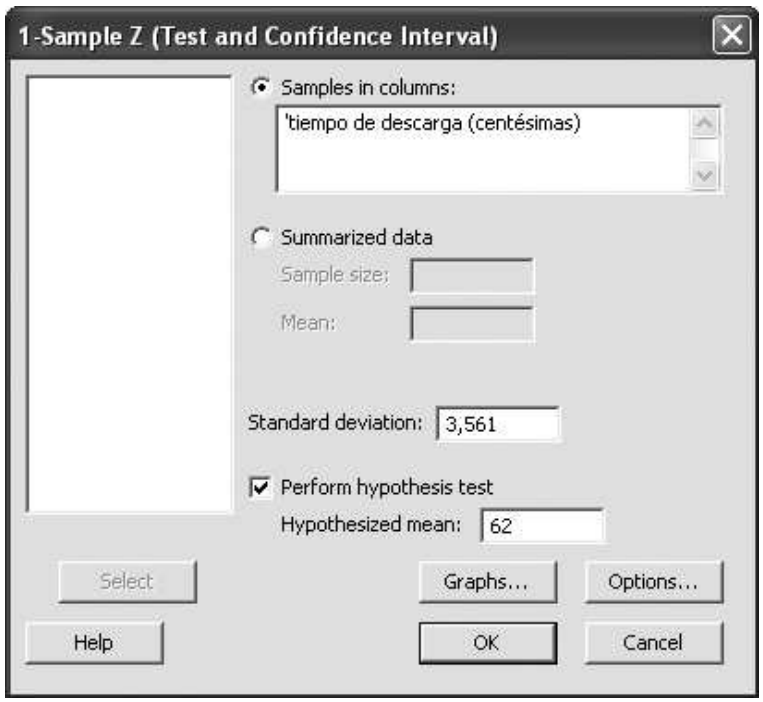
Figura 18. Gráfico de normalidad



Observando el *p*-valor se puede concluir que los datos siguen una distribución normal. Pudiendo asegurar que X sigue una distribución normal, la media muestral también sigue una distribución normal.

b) El contraste de hipótesis será H_0: $\mu = 62$ vs. H_1: $\mu \neq 62$. Es un contraste bilateral a un nivel de confianza de 0,90 (figura 19).

Figura 19. Pasos a seguir para realizar el contraste de hipótesis



Los resultados de Minitab son los que muestra la figura 20.

Figura 20. Resultados del contraste de hipótesis e intervalo de confianza del 90% (desviación típica población conocida)

One-Sample Z: tiempo de descarga (centésimas)

```
Test of mu = 62 vs not = 62
The assumed standard deviation = 3,561

Variable  N    Mean  StDev  SE Mean       90% CI          Z
tiempo  101  62,267  3,561    0,354  (61,685; 62,850)  0,75

Variable                      P
tiempo de descarga (cent.)  0,451
```

Se observa que el *p*-valor es 0,451, por lo tanto, como *p*-valor > $\alpha = 0{,}10$, no se puede rechazar la hipótesis nula, luego se acepta que el tiempo medio es de 62 centésimas por segundo.

c) El intervalo de confianza para el tiempo medio es (**61,685**; **62,850**), es coherente con los resultados esperados, ya que contiene al valor medio de 62 centésimas de segundo.

d) Análogamente se realiza el contraste de hipótesis para la media de la población con desviación típica desconocida, se selecciona ***Stat*** > ***Basic Statistic*** > ***1-Sample t***, obteniéndose los resultados de la figura 21.

Figura 21. Resultados del contraste de hipótesis e intervalo de confianza del 90% (desviación típica población desconocida)

One-Sample T: tiempo de descarga (centésimas)

```
Test of mu = 62 vs not = 62

Variable  N    Mean  StDev  SE Mean       90% CI          T
tiempo   101  62,267  3,561    0,354  (61,679; 62,856)  0,75

Variable                      P
tiempo de descarga (cent.)  0,452
```

El *p*-valor es 0,452 > 0,10, nos indica que se puede aceptar la hipótesis de que el tiempo medio es de 62 centésimas por segundo.

Continuando con el mismo ejemplo, se va a considerar que una página no es satisfactoria cuando tarde en ser descargada más de 68 centésimas. Los programadores afirman que el porcentaje de páginas para las que el tiempo de descarga no es satisfactorio no supera el 10%.

e) Se calculará un intervalo de confianza para la proporción de páginas no satisfactorias, a un nivel de confianza del 95%.

f) ¿Hay evidencias, al nivel 0,05, para rechazar la afirmación de los programadores? Se plantearán las hipótesis que se deben contrastar y se efectuará el contraste.

e) Para calcular el intervalo de confianza de la proporción de páginas no satisfactorias, a un nivel de confianza del 95%, se selecciona ***Stat*** > ***Basic Statistics*** >***1 Proportion*** (figura 22).

Observando la figura 23 de datos, se ve que únicamente hay 9 páginas que superan las 68 centésimas de segundo, o lo que es lo mismo, 9 páginas de las 101 se considera el tiempo de descarga no satisfactorio.

Figura 22. Pasos a seguir para obtener un intervalo de confianza del 95% para la proporción

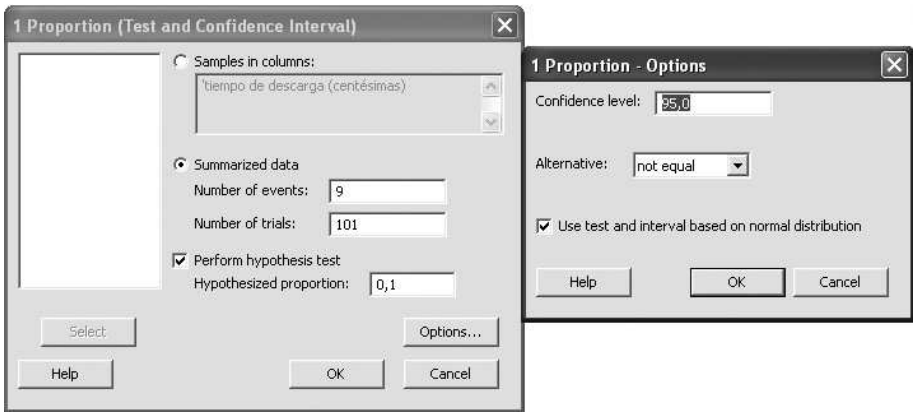


Figura 23. Resultados del intervalo de confianza del 95% para la proporción de páginas no satisfactorias

Test and CI for One Proportion

Test of p = 0,1 vs p not = 0,1

Sample	X	N	Sample p	95% CI	Z-Value	P-Value
1	9	101	0,089109	(0,033546; 0,144671)	-0,36	0,715

Using the normal approximation.

El intervalo de confianza obtenido con un nivel de confianza del 95% es **(0,033546; 0,144671)**.

f) Debemos plantear un contraste unilateral para la proporción de páginas no satisfactorias:

$\left.\begin{array}{l} H_0 : p = 0,1, \\ H_1 : p > 0,1, \end{array}\right\}$, donde p representa la proporción de páginas para las que el tiempo de descarga no es satisfactorio (figura 24).

Figura 24. Pasos a seguir para realizar el contraste de hipótesis

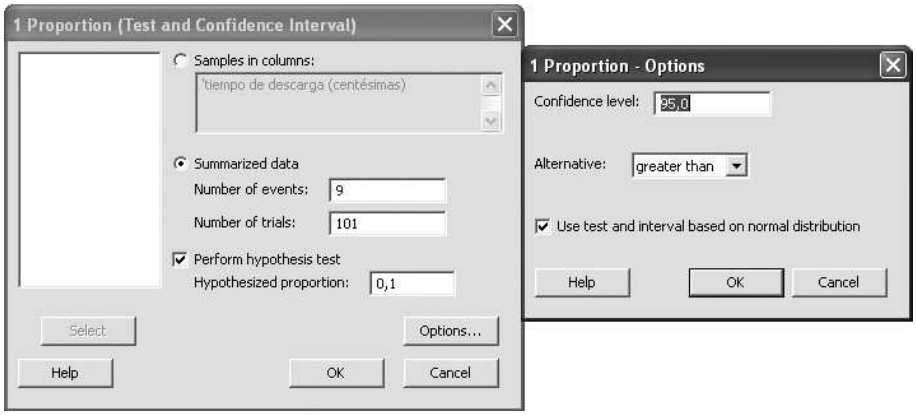


Figura 25. Resultados del contraste de hipótesis para la proporción de páginas

Test and CI for One Proportion

Test of p = 0,1 vs p > 0,1

Sample	X	N	Sample p	95% Lower Bound	Z-Value	P-Value
1	9	101	0,089109	0,042479	-0,36	0,642

Using the normal approximation.

Según se muestra en la figura 25, el *p*-valor del contraste vale lo siguiente: *p*-valor = 0,642. Como es mayor que 0,05, se acepta la hipótesis nula, luego se acepta la afirmación de los programadores de que el porcentaje de páginas no supera el 10%.

Resumen

En este módulo se presentan las distribuciones muestrales. Se analiza cómo seleccionar una muestra aleatoria simple, cómo se pueden emplear los datos obtenidos con ella para desarrollar estimaciones puntuales de los parámetros de población. La distribución de probabilidad de estas variables aleatorias se llama *distribución muestral*. En particular, se describen las distribuciones de la media de la muestra $\bar{x}$, de la proporción muestral $\hat{p}$ y de la varianza muestral s^2. Después de desarrollar las fórmulas de la desviación típica o error estándar para esos estimadores, se indica que el teorema central del límite es la base para usar una distribución normal de probabilidades y aproximar esas distribuciones muestrales en el caso de muestra grande.

Además, también se desarrollan estimaciones de intervalos de confianza de parámetros de una población. En este módulo se han utilizado la distribución Z normal estándar, la t de Student y la chi-cuadrado χ^2 para construir intervalos de confianza. Se determina el tamaño de muestra necesario para que los estimadores de intervalo de μ y de p tengan un nivel especificado de precisión.

Finalmente, en este módulo se ha presentado la metodología para realizar contrastes clásicos de hipótesis, comenzando con los argumentos para tomar decisiones en condiciones de incertidumbre. Las decisiones se toman rechazando una hipótesis nula si hay pruebas contundentes a favor de la hipótesis alternativa. Pueden cometerse dos tipos de error: un error de tipo I, que se comete cuando se rechaza la hipótesis nula, cuando es verdadera, y un error de tipo II, que se comete cuando no se rechaza la hipótesis nula, cuando no es verdadera, presentando diversos métodos y reglas de decisión específicos para realizar contrastes. La regla de rechazo para todos los procedimientos implica comparar el valor del estadístico con un valor crítico y también utilizando el p-valor para pruebas de hipótesis, la regla es rechazar la hipótesis nula siempre que el p-valor sea menor que α.

Ejercicios de autoevaluación

1) Una biblioteca presta un promedio de μ =320 libros por día, con desviación estándar σ = 75 libros. Se tiene una muestra de 30 días de funcionamiento, y $\bar{x}$ es la cantidad de la media de la muestra de libros prestados en un día.

a) Presente la distribución muestral de $\bar{x}$.
b) ¿Cuál es la distribución estándar de $\bar{x}$?
c) ¿Cuál es la probabilidad de que la media de una muestra de 30 días sea entre 300 y 400 libros?
d) ¿Cuál es la probabilidad de que la media de una muestra sea de 325 o más prestamos diariamente?

2) Un investigador informa los resultados de una encuesta diciendo que el error estándar de la media es de 20.La desviación estándar de la población es de 500.

a) ¿De qué tamaño fue la muestra que se usó en esta encuesta?
b) ¿Cuál es la probabilidad de que el error estimado quede a ± 25 o menos de la media de la población?

3) Cada curso escolar, una prestigiosa universidad oferta becas a sus estudiantes para ampliar estudios en el extranjero. De la experiencia recogida en anteriores convocatorias, se observa que las calificaciones medias de los expedientes aspirantes a obtener una beca se distribuyen según una normal de media 6,9 puntos y desviación estándar 0,7 puntos. Para entender la aplicación del teorema central del límite, generar con Minitab 50 muestras aleatorias de 100 observaciones cada una, que corresponden a la población normal anterior $N(6{,}9,\ 0{,}7)$.

a) Calcular en una nueva columna la media de las 50 muestras anteriores.
b) Comentar los resultados haciendo referencia al teorema central del límite.
c) Realiza el *dotplot* asociado a una de las muestras.
d) Compara estos resultados con la media de la población, y el valor de la desviación estándar de la media muestral con la desviación estándar de la población y explica la relación entre ambos valores.

4) Un estudio previo nos dice que el servicio de préstamo diario de libros de las bibliotecas de una ciudad sigue una distribución normal con una media de 300 ejemplares prestados, con una desviación estándar de 10. Una inspección quiere verificar si estos datos son correctos. Para hacerlo, coge una muestra de los préstamos diarios de 10 bibliotecas y obtiene una media de 285 ejemplares prestados.

a) ¿Cuál es la probabilidad de que si la media es verdaderamente de 300 ejemplares prestados se obtenga una media de préstamos igual o inferior a los 285 ejemplares en las 10 bibliotecas que componen la muestra?
b) Determinar un intervalo de confianza del 90% para la media de préstamos teniendo en cuenta los datos de la muestra.
c) ¿Qué decisión lógica debería tomar el inspector?

5) En la página web de una editorial aparecen dos números de teléfono. Hemos comprobado, después de analizar 400 llamadas del teléfono, que el intervalo entre llamadas tiene una varianza de 2.

Suponiendo normalidad, indicad si podemos considerar, a un nivel de confianza del 90%, que la varianza del intervalo entre llamadas del primer número es inferior a 1,7.

6) El responsable de comunicaciones de un centro de documentación afirma que la media del tiempo de transferencia de un fichero de tamaño 2Mb es superior a 30 segundos. Para comprobar esta afirmación se tomó una muestra de tiempos de transferencia de 12 ficheros de 2Mb, obteniendo que la media y la desviación estándar muestrales valen $\bar{x} = 30{,}2,\ s = 1{,}833$ (en segundos).

a) Suponiendo que el tiempo de transferencia se distribuye normalmente a partir de los datos muestrales obtenidos, ¿tenemos suficientes evidencias para aceptar la afirmación del responsable? (Tomad α = 0,05). Encontrad el *p*-valor del contraste.

Si además de disponer de estas observaciones nos hubiesen dado como información adicional (obtenida de experiencias previas) que la varianza del tiempo de transferencia es de $\sigma^2 = 9{,}2$ segundos^2, ¿hubiéramos llegado a la misma conclusión que en el apartado anterior? Encontrad el *p*-valor del contraste (Tomad α = 0,05).

Solucionario

1)

a) Normal con $\mu = 320$ y desviación típica 13,69
b) 13,69
c) 0,8558
d) 0,3557

2)

a) 625
b) 0,7888

3)

De esta manera obtenemos las 50 muestras con 100 observaciones cada una.

Worksheet 1 ***

↓	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
1	6,02255	7,51039	7,56941	6,73114	6,89504	7,27531	6,49352	7,07449	7,72782	8,01131	6,51824	7,31135	5,70901	7,18919	6,23128	7,986(
2	5,82796	7,23382	8,63014	7,67923	6,82915	7,50193	6,80251	6,83211	6,08951	7,08834	7,44647	7,36977	6,94125	6,82865	8,08839	6,073(
3	5,89460	6,19736	7,48944	7,16782	7,30042	6,23994	5,55889	5,92865	6,65836	7,20623	6,29604	6,50873	7,42008	6,63113	7,64693	7,277
4	7,91373	6,63874	6,88954	7,74717	7,75320	6,93165	6,16663	5,94975	7,84482	7,35052	6,62743	7,61428	6,31124	6,59273	6,66213	5,852(
5	6,96531	6,54096	7,83251	6,71421	5,76998	6,48673	6,47134	6,87821	7,57085	6,96200	7,26595	7,65586	7,88420	8,22074	7,28688	5,476(
6	6,38379	8,14155	5,26726	7,28067	6,89716	6,26790	7,12585	5,82249	6,81010	7,49986	6,99193	5,76598	7,32813	5,52918	7,05477	5,853
7	6,98679	7,50650	7,06138	7,30100	6,61602	7,20820	8,00421	6,49912	8,02260	7,28351	5,36631	6,93591	7,84171	6,61069	6,50505	6,263
8	6,28190	8,55299	8,20722	7,29348	6,51880	7,74509	6,95879	7,46706	7,55387	7,82200	7,22971	7,12365	7,03669	6,01653	8,28156	6,957(
9	5,85308	6,42670	7,24762	7,50398	6,46682	7,32013	6,42494	5,69820	6,13376	6,79857	7,15053	6,40466	6,38444	6,24852	6,05964	6,250

Project Man...

Current Worksheet: Worksheet 1 18:19

a) En la columna C101 se muestran las medias maestrales.

Worksheet 1 ***

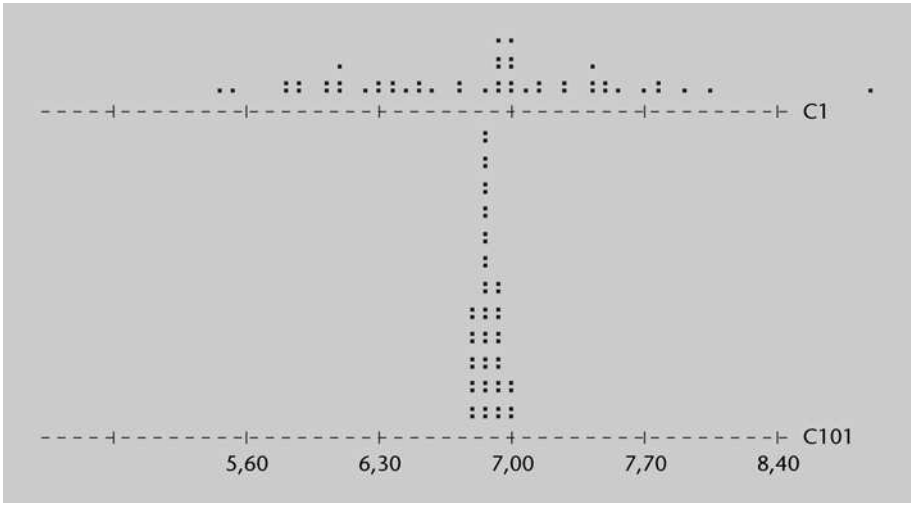
↓	C98	C99	C100	C101
1	5,49495	7,33352	6,49861	6,84032
2	7,27693	5,57569	5,13667	6,90402
3	6,86654	6,96175	7,56928	6,87066
4	7,85956	6,10293	7,09721	6,87309
5	7,68902	4,92515	7,13723	6,89323
6	6,51449	5,78576	7,18556	6,81035
7	7,46093	6,67470	5,89898	6,87570
8	5,41243	7,05719	7,60637	6,98587
9	8,66592	7,03988	7,55059	6,85254

Project Man...

Current Worksheet: Worksheet 1

b)

dotplot: C1; C101



c) Tras haber generado 50 muestras de datos provenientes de una distribución normal de media 6,9 y desviación estándar 0,7, observamos que el primer *dotplot* parece corresponder a una distribución normal.

Asimismo, el segundo *dotplot* corresponde a la distribución de las medias de las muestras y también corresponde a una distribución normal.

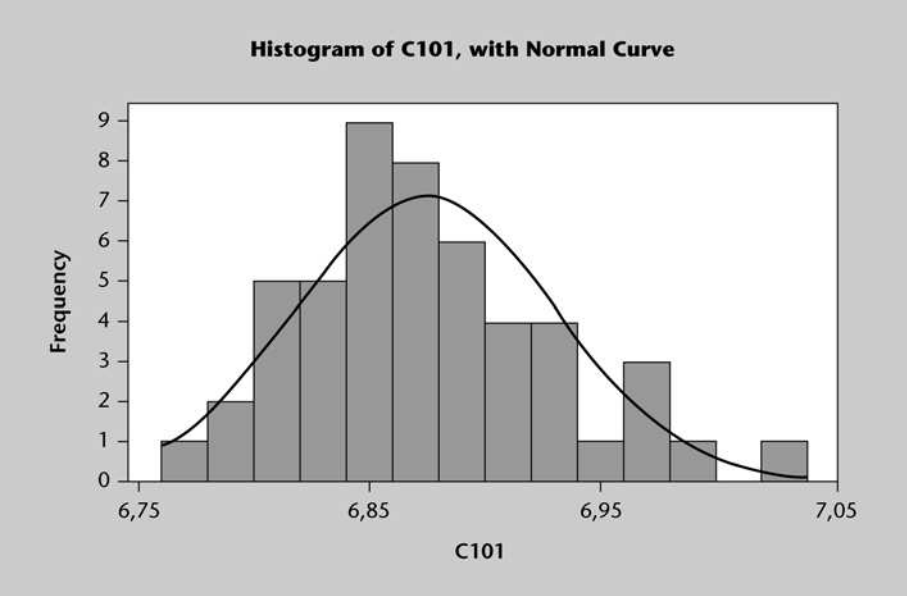
Esto indica que las medias de estas muestras siguen una distribución normal. Esta propiedad es la que enuncia el TCL, sea cual sea la distribución de los datos, la media muestral (con un tamaño de muestra *n* suficientemente grande) de una colección de datos sigue una distribución normal.

d) Estudiaremos la distribución de estas medias muestrales:

Descriptive Statistics: C101

Variable	N	Mean	Median	TrMean	StDev	SE Mean
C101	50	6,9035	6,8914	6,9016	0,0660	0,0093

Variable	Minimum	Maximum	Q1	Q3
C101	6,7837	7,0412	6,8507	6,9603



El histograma de frecuencias se aproxima a la curva normal, es simétrica.

La media muestral coincide con la media de la población, $\mu = \overline{x} = 6,9$.

La desviación estándar de la media muestral será aproximadamente el error estándar.

Si la variable tiene desviación estándar conocida *s* (en la población), el error estándar se puede calcular como:

$$\frac{\sigma}{\sqrt{n}}$$

Como consecuencia, podemos decir que la media muestral sigue una distribución normal $N(\mu, \frac{\sigma}{\sqrt{n}})$, que se puede aproximar a una $N(0,1)$, realizando un cambio de variable (tipificación): $Z = \frac{X - \mu}{\sigma / \sqrt{n}}$.

4)

a) Si estandarizamos la puntuación de 285, resulta un valor *z* de –4,74, lo que supone (mirando las tablas de la normal) aproximadamente que el 0% es la probabilidad de obtener dicha puntuación.

$$p(X < 285) = p\left(Z < \frac{285 - 300}{10/\sqrt{10}}\right) = p(Z < -4,74) \approx 0$$

b) El intervalo de confianza es:

$$z_{\alpha/2,} = 1{,}64$$

$$I = 285 \pm 1{,}64 \cdot \frac{10}{\sqrt{10}} = 285 \pm 5{,}17 \qquad [279{,}83;\ 290{,}19]$$

c) 300 está fuera del intervalo y, por lo tanto, con un nivel de confianza del 90%, podremos afirmar que la media no llega a 300 ejemplares, sino que está por debajo.

5) La hipótesis nula es $\sigma^2 = 1{,}7$ y la alternativa es $\sigma^2 < 1{,}7$.

El estadístico de contraste es: $\chi^2 = \frac{(400-1)s^2}{1{,}7}$, donde s^2 es la varianza muestral. Entonces $\chi^2 = 469{,}412$ y su distribución es la de χ^2 la con $400-1 = 399$ grados de libertad.

En este caso, el *p*-valor vale $P(\chi^2 < 469{,}412) = 0{,}991406$ y por lo tanto, no rechazamos la hipótesis nula: no podemos afirmar que sea inferior a 1,7. El valor crítico es 363,253.

6)

a) Hemos de hacer el contraste de una media con varianza desconocida. Las hipótesis nula y alternativa son: $\begin{matrix} H_0 : \mu = 30 \\ H_1 : \mu > 30 \end{matrix}$, donde μ representa la media del tiempo de transferencia de un fichero de 2Mb. El estadístico de contraste es $t = 0{,}378$. El valor crítico valdrá: $t_{0.05,11} = 1{,}80$. Como que $t < t_{0.05,11}$, aceptamos la hipótesis nula y concluimos que la afirmación del responsable es cierta. Si quisiéramos hallar el *p*-valor, éste sería: $p = p(t_{11} > 0{,}378) \approx 0{,}36$. Como es un *p*-valor alto, mayor que 0,05, aceptamos la hipótesis nula tal y como hemos hecho antes.

b) Hemos de hacer el contraste de una media con varianza conocida.

La hipótesis nula y la alternativa son: $\begin{matrix} H_0 : \mu = 30 \\ H_1 : \mu > 30 \end{matrix}$, donde μ representa la media del tiempo de transferencia de un fichero de 2Mb.

El estadístico de contraste es: $z = \frac{\bar{x} - 30}{\sigma/\sqrt{12}}$, donde $\bar{x}$ es la media muestral y σ es la desviación estándar poblacional. La distribución de *z* es la de una normal $N(0,1)$. La media y la desviación estándar poblacionales valen respectivamente: $\bar{x} = 30{,}2$, $\sigma = \sqrt{9{,}2} \approx 3{,}03$. El valor del estadístico de contraste es: $z \approx 0{,}228$.

El valor crítico valdrá: $z_{0{,}05} \approx 1{,}645$. Como $z < z_{0{,}05}$, volvemos a aceptar la hipótesis nula y concluimos que la afirmación del responsable no es cierta. Si quisiéramos hallar el *p*-valor, éste sería: $p = p(z > 0{,}228) \approx 0{,}41$. Como es un *p*-valor alto, mayor que 0,05, aceptamos la hipótesis nula como hemos hecho anteriormente. Por tanto, hemos llegado a la misma conclusión que en el apartado anterior.